Chapter 5
Recognition of Induced Sentiment when Reviewing Personal Egocentric Photos

Abstract

Lifeblogging is a process of collecting a rich source of information about the daily life of people. The availability and use of egocentric data are rapidly increasing due to the growing use of wearable cameras. In this work, we introduce the problem of sentiment analysis in egocentric events focusing on the moments that compose the images recalling positive, neutral or negative feelings to the observer. Given egocentric photostreams capturing the wearer’s days, we propose a method for the classification of the sentiments in egocentric pictures based on global and semantic image features extracted by Convolutional Neural Networks. Such moments can be candidates to retrieve according to their possibility of representing a positive experience for the camera’s wearer. We carried out experiments on an egocentric dataset, which we organized in 3 classes on the basis of the sentiment that is recalled to the user (positive, negative or neutral). Our model makes a step forward opening the door to sentiment recognition in egocentric photostreams.
5. Recognition of Induced Sentiment when Reviewing Personal Egocentric Photos

5.1 Introduction

Mental imagery is the process in which a feeling of an experience is invoked by a person in the absence of external stimuli. Therapists assume that it is directly related with emotions (Holmes and et al., 2006), leading to some questions about the effect of images that depict past moments: Can an image make the process of mental imagery easier? or Can specific images help us to invoke feelings and moods?

Although our mood is influenced by the environment and social context that surrounds us, egocentric data do not always catch our attention or induce the same emotion. We consider that the creation of an electronic diary of positive moments will help to improve the perception of the user of his/her own life. Usually, users are interested in keeping special moments, images with sentiments that will allow them in the future to re-live the personal moments captured by the camera. An automatic tool for sentiment analysis of egocentric images is of high interest to make possible the processing of the big collection of lifelogging data and keeping out just the images of interest i.e. of high charge of positive sentiments.

We approach this problem in this work from two different perspectives. On one hand, we propose to analyse the relation of semantic concepts extracted from images that belong to the same scene. To this end, we defined a classification model where one-vs-all SVM classifiers were trained and evaluated with the features describing semantic and global information from images. On the other hand, we propose to combine semantic concepts, given that they have associated sentiment values (Borth et al., 2013), with general visual features extracted by a CNN (Krizhevsky, Sulskever and Hinton, 2012). Semantic concepts extracted from images represent a finite subset of what is present in the image, not covering the whole image content. Visual features extracted by CNNs can help to summarize the whole image content at an intermediate level.

Our contribution is an analytic tool for positive emotion retrieval seeking events that best represent a pleasant moment to be invoked within the whole set of a day photo-stream. We focus on the event’s sentiment description where we are observers without inner information about the event, i.e. from an objective point of view of the moment under analysis.

5.2 Related works

Automatic sentiment image analysis is a complicated task since there is no consensus between the different sentiment ontologies presented in the literature. Table 5.1 illustrates the ambiguity of the problem, reporting several sentiments ontology related to images. The first group (Machajdik and Hanbury, 2010; You et al., 2016; Yi

---
5.2. Related works

et al., 2014) assigns 8 main sentiments as excitement, awe or sadness to the images with assigned discrete positive (1) and negative (-1) sentiment value. The second group (Dan-Glauser and Scherer, 2011; Lang et al., 1997) defines a different set of sentiments as valence or arousal and discrete positive (1), neutral (0), or negative (-1) values assigned to the images according to the sentiments. In contrast, the third group (Nojavanasghar and et al., 2016) assigns up to 17 sentiments (6 basics and 9 complex) and each image of the dataset is assigned a continuous value in a scale from 1 to 4. Given the ambiguity of the semantic sentiment assignment, with labels difficult to classify into positive or negative sentiments, the last group (Borth et al., 2013) defines up to 3244 Adjective Noun Pairs (ANP) (e.g. ‘beautiful girl’) and assigns a continuous sentiment value in a range of [-2,2] to them. The main idea is that the same object according to its appearance has positive or negative sentiment value like ‘angry dog' (-1.55) and ‘adorable dog' (+1.45). A natural question is until which extent the 3244 ANPs represent a scene captured by the image, taking into account the difficulty to detect them automatically (Mean average accuracy ∼25%).

<table>
<thead>
<tr>
<th>DataSets</th>
<th>Source</th>
<th>Images</th>
<th>Semantic sentiment labels</th>
<th>Sentiment Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract &amp; Artphoto &amp; Machajdik and Hanbury, 2010</td>
<td>Flickr &amp; Instagram</td>
<td>280 &amp; 806</td>
<td>positive: contentment, amusement, excitement, awe, negative: sadness, fear, disgust, and anger</td>
<td>{1,-1}</td>
</tr>
<tr>
<td>You’s Dataset &amp; You et al., 2016</td>
<td>Flickr</td>
<td>23000</td>
<td>positive: contentment, amusement, excitement, awe, negative: sadness, fear, disgust, and anger</td>
<td>{1,-1}</td>
</tr>
<tr>
<td>CASIA-WebFace &amp; Yi et al., 2014</td>
<td>Flickr &amp; Instagram</td>
<td>494k</td>
<td>happy, neutral, sad, surprise</td>
<td>{1,0,-1}</td>
</tr>
<tr>
<td>IAPS &amp; Lang et al., 1997</td>
<td>Flickr</td>
<td>1182</td>
<td>valence, arousal, and dominance</td>
<td>{1,2}</td>
</tr>
<tr>
<td>GAPED &amp; Dan-Glauser and Scherer, 2011</td>
<td>Flickr &amp; Instagram</td>
<td>732</td>
<td>valence, arousal, and normative significance</td>
<td>{1,0,-1}</td>
</tr>
<tr>
<td>EmoReact &amp; Nojavanasghar and et al., 2016</td>
<td>Youtube</td>
<td>1102 clips</td>
<td>17 sentiments: 6 basic emotions: happiness, surprise, negative: sadness, fear, disgust, and anger, and 9 complex emotions: curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment, frustration.</td>
<td>{1,4}</td>
</tr>
<tr>
<td>VSO + &amp; TwitterIm &amp; Borth et al., 2013</td>
<td>Twitter</td>
<td>0.5M</td>
<td>Not, but Adjective Noun Pairs (3244)</td>
<td>Flickr{-2,2} Twitter{-1,1}</td>
</tr>
<tr>
<td>You &amp; Neubert &amp; You and et al., 2015</td>
<td>Twitter</td>
<td>1269</td>
<td>Non-semantic labels: Positive and Negative</td>
<td>(1,-1)</td>
</tr>
<tr>
<td>UBRCU &amp; EgoSenti*</td>
<td>Wearable Camera</td>
<td>12088</td>
<td>Non-semantic labels: Positive, Neutral and Negative</td>
<td>(1,0,-1)</td>
</tr>
</tbody>
</table>

Table 5.1: Different image sentiment ontologies.

Given the difficulty of image sentiment determination, ambiguity and lack of consensus in the bibliography, added by the difficulty given by the egocentric im-
ages, we focus on the image sentiment as a discrete value expressing a ternary sentiment value (positive \((1)\), negative \((-1)\) or neutral \((0)\) value) similar to (You and et al., 2015). Egocentric data is of special difficulty, since we do not observe the wearer and his/her, i.e. from facial or corporal expressions, but rather from the perspective of what the user sees. Moreover, in real life, fortunately, negative emotions have much less prevalence than neutral and positive, that makes very difficult to have enough examples of negative egocentric images and events. Thus, the problem we address in this article is what effect an egocentric image or event has on an observer (positive, neutral or negative) (see Fig.5.1), instead of attempting to specify an explicit semantic image sentiment like sadness; and how to develop automatic tool for sentiment value detection (positive, vs. neutral vs. negative) and egocentric dataset in order to validate its results. Going further, in contrast to the published work, we claim to automatically analyse the sentiment value of egocentric events i.e. a group of sequential images that represents the same scene. In the case of egocentric images, the probability that a single image describes an event is low; there are a lot of images that just capture wall, sky, ground or partially objects. For this reason, we are interested to automatically discover how the event captured by the camera influences the observer, that is to automatically determine the ternary sentiment values of the events, which are richer in information and involve the whole moment’s experience. For example, an event being in a dark and narrow, grey space would influence negatively, a routine scene like working in the wearer’s office could influence the observer neutrally and an event where the wearer has spent some time with friends in a nice outdoor space could influence positively to the observer.

Automatic sentiment analysis from images is a recent research field. In the literature, sentiment recognition in conventional images has been approached by computing and combining visual, textual, or audio features (Nojavanasghar and et al., 2016; Poria and et al., 2014; Wang et al., 2014; You and Et, 2016). Other characteristics, such as facial expressions have also been used for sentiment prediction (Yuan and et al., 2013). The combination of visual and textual features extracted from images is possible due to the wide use of online social media and microblogs, where images are posted accompanied by short comments. Therefore, multimodal approaches were proposed, where both sources of information are merged (Wang et al., 2014; You and Et, 2016) for automatic sentiment value detection.

Recently, with the outstanding performance of Convolutional Neural Networks (CNN), several approaches to sentiment analysis relied on deep learning techniques for classification and/or features extraction combined with other networks or methods (Campos and et al., 2015; Levi and Hassner, 2015; You et al., 2016; Yu et al., 2016). The work in (You et al., 2016) applies fine-tuning on the AlexNet to classify the 8 emotions: sadness, angry, content, etc. In contrast, in (Campos and et al., 2015)
they propose to fine-tune CaffeNet with oversampling to classify into Positive or Negative sentiments. In (Levi and Hassner, 2015) a novel transformations of image intensities to 3D spaces is proposed to reduce the amount of data required to effectively train deep CNN models. In (Yu et al., 2016) the authors use logistic regression to classify into 3 sentiments using CNN features. In (Chen et al., 2014), the authors perform a fine-tuning on a CNN model and modify the last layer to classify 2089 ANPs. However, no work has addressed the sentiment image and event analysis in egocentric datasets.

5.3 Sentiment detection by global features analysis

In this section, we describe the proposed method for sentiment recognition from egocentric photo-streams, which is based on visual (extracted by CNN) and semantic (in terms of ANPs) features extracted from the images. An architectural overview of the proposed system is depicted in Fig. 5.2.

a) Temporal Segmentation:

Given that egocentric images have a smaller field of view and thus do not capture entirely the context of the event, we need to detect the events of the days. To this aim, we apply the SR-Clustering algorithm for temporal segmentation of photo-streams (Dimiccoli et al., 2017). The clustering procedure is performed on an image representation that combines visual features extracted by a CNN with semantic features in terms of visual concepts extracted by Imagga’s auto-tagging technology (http://www.imagga.com/solutions/auto-tagging.html).

b) Features Extraction:

For the computation of the semantic features in terms of the ANPs, we use the DeepSentiBank Network (Chen et al., 2014). Given an image, the DeepSentiBank network considers the 2089 best performing ANPs. Applying the DeepSentiBank on them gives a 2089-D feature vector, where the feature values correspond to the ANPs likelihood in the image. These values are multiplied by the sentiment value associated with the concepts. Note that each ANP has a positive or negative sentiment value assigned, but not 0 for neutral sentiment.

However, the 2089 ANPs not necessarily have the power to explain the “richness” of any scene in an image. Hence, we integrate the ANPs feature vector with a feature descriptor provided by the penultimate layer if a CNN (Krizhevsky, Sulskever and Hinton, 2012) that summarizes the whole context of the image. The resulting
feature vector is composed of 4096 features. We combine the ANPs and the CNN feature vectors into a 6185-D feature vector, in order to construct a more reliable and rich image representation that relates image semantics expressed by the ANPs with clear sentiment value with the CNN cues as an intermediate image representation. We apply the Signed Root Normalization (SRN) to transform the CNN feature vectors to a more uniformly distributed space followed by a $l_2$-normalization (Zheng et al., 2014).

c) Classification:

We use the proposed feature vectors to train a multi-class SVM classifier due to its high generalization capability (Joachims, 2000). This is ensured by the SVM learning algorithm that finds a separation hyperplane that maximizes the separation margin.
5.3. Sentiment detection by global features analysis

Combination of events

SVM1
SVM2
SVM3
Multiclass SVM

Features label
Global + Semantic

Photo Stream
...
SR-Clustering
Features Extraction
...
...
...
...
...
(a) (b) (c) (d)

Figure 5.2: Architecture of the proposed method. (a) Temporal segmentation of the photo-stream into events. (b) CNN and ANPs features are extracted from the images and (c) used as input to the trained multi-class SVM model. (d) The model labels the input image as Positive, Neutral or Negative.

between the classes. We employ a 1-vs-all design for the multi-class problem, as suggested in (Foggia et al., 2015). The cardinality of the classes in the proposed dataset is not balanced, which affects the computation of the training error cost on the corresponding positive and negative samples. We set the cost of the training error on the positive and negative class according to their cardinality for each SVM of the pool of classifiers. In the implementation of the SMVs, we set the training error costs according to the ratio \( r = n^-/n^+ \), where \( n^+ \) and \( n^- \) are the number of positive and negative examples, respectively. At this stage, the decision of the classifier is taken at image level. To classify an event, we use a majority vote on the image level classification output.

5.3.1 Experimental Setup

Data set

We collected a dataset of 12471 egocentric pictures, which we call UBRUG-EgoSenti. The users were asked to wear a Narrative Clip Camera, which takes a picture every 30 seconds, hence each day around 1500 images are collected for processing. The images have a resolution of 5MP and JPG format.

We organize the images into events according to the output of the SR-clustering algorithm (Dimiccoli et al., 2017). From the originally recorded data, we discarded those events that are composed of less than 6 images, so obtaining a dataset composed of 12088 images grouped in a total of 233 events, with an average of 51.87
5. Recognition of Induced Sentiment when Reviewing Personal Egocentric Photos

images per event and std of 52.19. We manually labelled the events following how the user felt while reviewing them by assigning Positive, Negative and Neutral values to them, some examples of which are given in Fig 5.1. The dataset, for which the details are in Table 5.2 is publicly available and can be downloaded from: http://www.ub.edu/cvub/dataset/.

<table>
<thead>
<tr>
<th>Class</th>
<th>Images</th>
<th>#Events</th>
<th>Mean Im Event</th>
<th>Std Im Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>4737</td>
<td>83</td>
<td>57.07</td>
<td>52.34</td>
</tr>
<tr>
<td>Neutral</td>
<td>6169</td>
<td>107</td>
<td>57.65</td>
<td>57.18</td>
</tr>
<tr>
<td>Negative</td>
<td>1182</td>
<td>43</td>
<td>27.49</td>
<td>26.44</td>
</tr>
<tr>
<td>Total</td>
<td>12088</td>
<td>233</td>
<td>51.88</td>
<td>52.19</td>
</tr>
</tbody>
</table>

Table 5.2: Description of the UBRUG-EgoSenti dataset.

Experiments and Results

We carried out 10-fold cross-validation. Events from different classes are uniformly distributed among the various folds, which are thus independent of each other. We evaluated the performance of the proposed system on single images and at event level. For the UBRUG-Senti dataset, the ground truth labels are given at event level. All the images that compose a certain event, are considered as having the same label of such event. Given an event composed of $M$ images, we aggregate the $M$ classification decisions by majority vote. We measure the performance results of our method by computing the average accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Image Classification</th>
<th>Event Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pos</td>
<td>Neg</td>
</tr>
<tr>
<td>Semantic Features</td>
<td>mean</td>
<td>59.2</td>
</tr>
<tr>
<td>CNN Features</td>
<td>mean</td>
<td>70</td>
</tr>
<tr>
<td>Semantic+CNN Features</td>
<td>mean</td>
<td>72</td>
</tr>
</tbody>
</table>

Table 5.3: Performance results achieved at image and event level.

In Table 5.3 we report the results achieved by the proposed methods at image and event level. We achieved an average image classification rate of 59.60% with a standard deviation of 23.17, when we apply the proposed method. The average event classification rate is 68%, when the proposed features are employed, which corresponds to 82%, 73.5% and 49% for positive, negative and neutral events, respectively. Up to our knowledge, unfortunately, there is no work in the literature on egocentric image sentiment recognition neither event sentiment recognition to compared with. Even the works on image sentiment analysis in conventional images
5.4. Sentiment detection by semantic concepts analysis

(Campos and et al., 2015; Levi and Hassner, 2015; You et al., 2016; Yu et al., 2016) use different datasets and objectives (8 semantic sentiments vs. binary or ternary sentiment values) that make difficult their direct comparison. Fig. 5.3 shows some example results. As can be seen, the algorithm learns to classify events with the presence of routine objects into neutral events. Events wrongly classified as neutral are shown in Fig. 5.3(left) and Fig. 5.3(middle). As an example, the last row of Fig. 5.3(left) is classified as neutral, probably due to the presence of the pc in the image, while it was manually labelled as positive, because it shows social interactions. As for Fig. 5.3(left) and Fig. 5.3(right), events were mislabelled as negative probably due to the “homogeneity” and “greyness” of the images within the events, e.g. events were considered as negative when most of the information in the image corresponded to the asphalt of the road.

![Figure 5.3: Examples of the automatic event sentiment classification. The events are grouped based on the sentiment defined by the user: (right) Positive, (middle) Negative, and (left) Neutral. The events frame colour corresponds to the label given by the model: Positive (green), Negative (red) and Neutral (yellow).](image)

5.4 Sentiment detection by semantic concepts analysis

Given an egocentric photo-stream, we propose scene emotion analysis seeking for events that represent and can retrieve a positive feeling from the user. We apply event-based analysis since single egocentric images cannot capture the whole essence of the situation. By combining information from several images that represent the same scene, we get closer to a better understanding of the event.

a) Temporal segmentation:

We apply temporal segmentation on the egocentric photo-streams using the proposed method in (Dimiccoli et al., 2017). The clustering procedure is performed on an image representation that combines visual features extracted by a CNN with
5. Recognition of Induced Sentiment when Reviewing Personal Egocentric Photos

Figure 5.4: Sketch of the proposed method. First, a temporal segmentation is applied over the egocentric photo-stream (a). Later, semantic concepts are extracted from the images using the DeepSentiBank (Chen et al., 2014) (b). The semantic concepts with higher occurrence are selected as event descriptors (c). Finally, the ternary output is obtained by merging the sentiment values associated to the event’s semantic concepts (d).

semantic features in terms of visual concepts extracted by Imagga’s auto-tagging technology. In Fig. 5.3 we present some examples of events extracted from the dataset, we introduce below.

b) Event’s sentiment recognition:

The model relies on semantic concepts extracted from the images to infer the event sentiment associated. However, it relies not only on the semantic concepts extracted by the net with their associated sentiment, but also on how those semantic concepts can be interpreted by the user. We apply the DeepSentiBank Convolutional Neural Network (Chen et al., 2014) to extract the images semantic information since it is the only introduced model that extract semantic concepts (ANPs) with sentiment values associated. Given an image, the output of the network is a 2089-D feature vector, where the values correspond to the ANPs likelihood in the image.

Besides taking into account the sentiment associated with the ANPs, the influence of the common concepts within an event are also analysed. We categorize the noun into Positive, Neutral or Negative. There is a wide range of semantic concepts within the ontology, but many of them seem to repeat concepts that even from the

user perspective would be difficult to differentiate when looking at an image; such as ‘girl’ from ‘woman’ or ‘lady’.

When facing our egocentric images challenge, the VSO presents several drawbacks. On one hand, this tool is trained to recognize up to 2089 concepts, which cannot describe all possible scenarios. On the other hand, despite including that big amount of concepts, many of them categorize objects into categories difficult to visually interpret or differ by the human eye. Examples can be the distinction between ‘child’, ‘children’, ‘boy’, or ‘kid’ from an image. To overcome this problem, we generate a parallel ontology with what we consider an egocentric view of the concepts, i.e., we cluster the concepts a person would merge based on their semantic.

**Egocentric analysis of the VSO:** We cluster the semantic concepts based on the similarities between the noun components of the ANPs, which are computed using the wordNet tool\(^2\). Following what would be considered as similar from an egocentric point of view, we manually refine the resulted clusters into 44 categories. We label the clusters as Positive, Neutral or Negative. In Table 5.4, we present some of the ego-semantic clusters.

<table>
<thead>
<tr>
<th>Positive</th>
<th>Neutral</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>petals, rose</td>
<td>car, study</td>
<td>tumb, bug, nightmare</td>
</tr>
<tr>
<td>christmas, award</td>
<td>science, book</td>
<td>monument, insect</td>
</tr>
<tr>
<td>flora, snow</td>
<td>machine, history</td>
<td>grave, worm, noise</td>
</tr>
<tr>
<td>park, present</td>
<td>vehicle, economy</td>
<td>memory, cockroach</td>
</tr>
<tr>
<td>yard, santa</td>
<td>rally, market</td>
<td>memorial, cockroach</td>
</tr>
<tr>
<td>plant, cartoon</td>
<td>train, industry</td>
<td>stone, decay</td>
</tr>
<tr>
<td>garden, drawing</td>
<td>competition</td>
<td>graveyard, garbage</td>
</tr>
<tr>
<td>comics, toy</td>
<td>sculpture</td>
<td>cemetery, trash</td>
</tr>
<tr>
<td>illustration, toys</td>
<td>control, museum</td>
<td>grief, shit</td>
</tr>
<tr>
<td>humor, lego</td>
<td>statue, paper</td>
<td>pain</td>
</tr>
<tr>
<td></td>
<td>paper</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.4: Examples of clustered concepts based on their semantic similarity, initially grouped following the distance computed by the WordNet tool.

### 5.4.1 Sentiment Model

Given an event, the event’s sentiment analysis model (see Fig. 5.4) performs as follows:

1. Given the ego-photo-stream we apply the temporal segmentation, analyze events with a minimum of 6 images, i.e. that last for at least 3 minutes.

2. Extract the ANPs of each event frame and rank them by their probability \(Prob_{ANP}\) of describing an image.

---

\(^2\)http://wordnet.princeton.edu
3. Select the top-5 ANPs per image, since we consider that those are the concepts with higher relevance, thus better capturing the image’s information. After this step, the model ends up with a total of $M$ semantic concepts per event, where $M = \text{Number of images} \times 5$.

4. Cluster the $M$ semantic concepts based on their Wordnet-based nouns semantic distances. As a result, we have clusters of concepts with semantic similarity. For the event sentiment computation ($S_{\text{event}}$), focus on the largest cluster.

5. Finally, fuse the sentiment associated with the ANPs and noun’s cluster following the eq. (5.1):

$$S_{\text{event}} = \sum_j (\alpha \ast S_{\text{ANP}_j} + \beta \ast S_{\text{Noun}_j}), j = 1 : N_{\text{ANP}},$$

where $S_{\text{ANP}_j} = (S_{\text{VSO}_{\text{ANP}_j}} \ast \text{Prob}_{\text{ANP}_j})$, $S_{\text{VSO}_{\text{ANP}_j}}$ is the ANP’s sentiment given by the VSO and $S_{\text{Noun}_j}$ is the label of the noun, $\alpha$ and $\beta$ are the contributions (%) of the ANPs and the nouns. Take into account the probability associated to the ANPs aiming to penalize the ANPs with low relation to the image content.

5.4.2 Experimental Setup

Data set

We collected a dataset of 4495 egocentric pictures, which we call UBRUG-Senti. The user was asked to wear the Narrative Clip Camera fixed to his/her chest during several hours every day and was asked to continue with his/her normal life. Since the camera is attached to the chest, the frames vary following the user’s movement and describe the user’s view of his/her daily indoor/outdoor activities. It involves challenging backgrounds due to the scene variation, handled objects appearing and disappearing during images sequences, and the movement of the user. The camera takes a picture every 30 seconds, hence each day around 1500 images are collected for processing. The images have a resolution of 5MP and JPG format.

After the temporal clustering [Dimiccoli et al. (2017)], we obtained a dataset composed of 4495 images grouped in a total of 98 events. The events were manually labelled based on how the user felt while reviewing them. The labels assigned were Positive (36), Negative (43) and Neutral (19). Some examples are given in Fig 5.3.

3http://getnarrative.com/
5.4. Sentiment detection by semantic concepts analysis

Experiments and Results

During the experimental phase, we evaluated the contributions of ANPs and nouns by defining different combinations of \( \alpha \) and \( \beta \). We performed a balanced 5-fold cross-validation. For each of the folds, we used 80% of the total of events per label of our dataset and compute the best pair of \( \alpha \) and \( \beta \) values. This is a parameter selection process that is later re-evaluated in a test phase with a different set of events.

Validation: To evaluate the effectiveness of the scene detection approach, we use the Accuracy, as the rate of correct results, and the F-Score (F1). The F1 is defined as: 
\[
F_1 = \frac{2(RP)}{(R + P)},
\]
where \( P \) is the precision \((P = TP/(TP + FP))\), \( R \) is the recall \((R = TP/(TP + FN))\) and \( TP, FP \) and \( FN \) respectively are the number of true positives, false positives and false negatives of the event’s sentiment label correctly identified.

Results: Tables 5.5 and 5.6 present the results achieved by the proposed method at image and event level, respectively. The model achieves an average training accuracy of 73±3.8% and F-score of 59±5.4% and test accuracy of 75±8.2% and F-score of 61±13.2%, when \( \alpha = 0.8 \) and \( \beta = 0.2 \), i.e. when the ANP information is considered; although the major contribution comes from the noun sentiment associated. As expected, neutral events are the most challenging ones to classify.

<table>
<thead>
<tr>
<th>( \alpha = 0.8 )</th>
<th>( \alpha = 0.5 )</th>
<th>( \alpha = 0.2 )</th>
<th>( \beta = 0.8 )</th>
<th>( \beta = 0.5 )</th>
<th>( \beta = 0.2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours ( \beta = 0.2 )</td>
<td>0.60</td>
<td>0.63</td>
<td>0.73</td>
<td>0.35</td>
<td>0.43</td>
</tr>
<tr>
<td>Evaluating with weights ( \beta = 0.2 )</td>
<td>0.65</td>
<td>0.65</td>
<td>0.66</td>
<td>0.41</td>
<td>0.43</td>
</tr>
</tbody>
</table>

Table 5.5: Parameter-selection results

In order to contextualize our results, we fine-tune the well-known GoogleNet deep convolutional neural network [Ma et al., 2016] to classify into Positive, Neutral and Negative. We use 80%, 10% and 10% of the dataset for training, validation and testing respectively. The network achieves an accuracy of 55%.

From the results we can conclude that the application of the DeepSentiBank presents drawbacks when applied to egocentric photo-streams. To begin with and as commented before, the 2089 ANPs not necessarily have the power to represent what the image captured about the scene, taking into account the difficulty to detect them automatically (Mean average accuracy of the net ∼25%). Moreover, the ANPs
5. Recognition of Induced Sentiment when Reviewing Personal Egocentric Photos

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ours</strong></td>
<td>0.75±0.08</td>
<td>0.60±0.13</td>
</tr>
<tr>
<td>Evaluating</td>
<td>0.69±0.1</td>
<td>0.50±0.15</td>
</tr>
<tr>
<td>3 Clusters</td>
<td>0.74±0.1</td>
<td>0.58±0.15</td>
</tr>
<tr>
<td>Evaluating</td>
<td></td>
<td></td>
</tr>
<tr>
<td>with weights</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.6: Test set results

present the limitation that they are classified strictly into Negative or Positive concepts. Thus, moments from our daily routine, which are often considered as neutral, are difficult to recognize.

Sentiments recognition from an image or a collection of images is a difficult process due to its ambiguity. A challenge in the model construction for sentiment recognition consists in taking into account the bias due to the subjective interpretation of images by different users. Furthermore, the boundaries between neutral/positive and neutral/negative sentiments are not clearly defined. A neutral feeling is difficult to interpret. From the results, we observe that neutral events are the most challenging ones to classify. Another challenging aspect concerns the grouping of image sentiments into event sentiment, since events can have non-uniform sentiments.

A further step towards better understanding of the image and sentiment analysis is needed, due to the subjectivity of what an image can recall to different persons. To this aim, having annotations by different persons is critical to evaluate the inter- and intra-observer variability.

From the results, the intuition that we get is that non-routine events and specially when moments are social, have a higher probability of being positive. In contrast, routine events will most probably be considered as neutral. Negative events as accidents have low prevalence to be learned. Yet, hostile and empty environments could lead to negative sentiments too. Future works will address the study of emotional events and their relation to daily routine.

5.5 Discussion and conclusions

In this work, we proposed for the first time models and a dataset for sentiment recognition from egocentric images and events, recorded by a wearable camera.

Sentiment recognition from an image or a collection of images is a difficult process due to the subjectivity of the task. A challenge in the model construction for
sentiment recognition consists in taking into account the bias due to the subjective interpretation of images by different users. Furthermore, the boundaries between neutral/positive and neutral/negative sentiments are not clearly defined. A neutral feeling is difficult to interpret. From the results, we observe that neutral events are the most challenging ones to classify. Another challenging aspect is the fact that events are represented by groups of images that do not necessarily share the same associated sentiment. Thus, by giving a sentiment label to an event, we extrapolate it to the images that compose it, being aware that this might imply some errors.

In [Talavera, Radeva and Petkov, 2017] we first introduced a labelled dataset composed of 98 events. Later, in [Talavera, Strisciuglio, Petkov and Radeva, 2017] we extended it to 233 events, grouping 12088 images, from 20 days recorded by 3 different users.

The first proposed approach is based on the extraction of CNN and semantic features with associated sentiment value. It analyses semantic concepts called Adjective-Noun-Pairs (ANPs) extracted from the images, which have an associated sentiment value and describe the appearance of concepts in the images. The sentiment prediction tool is based on new semantic distance of ANPs and fusion of ANPs and nouns sentiments extracted from egocentric photo-streams. This model obtained a classification accuracy of 75% on the test set, with a deviation of 8% over the first version of the dataset. The second proposed approach is based on a classification model where one-vs-all SVM classifiers were trained and evaluated with the features describing semantic and global information from the images. Using the proposed method we obtained average events and image sentiment accuracy of 68.17% and 58.60%, with a standard deviation of 30.07% and 23.17%, respectively.

Analysing the obtained results, we conclude that the polarity of the ANPs makes it difficult to classify 'Neutral' events. However, most of our daily life is composed of such events, which can be considered as routine. Furthermore, we get the intuition that non-routine events have a higher probability of being positive, especially when moments are social. In contrast, routine events will most probably be considered as neutral. Negative events as accidents have low prevalence to be learned. Yet, hostile and empty environments could lead to negative sentiments too. Future works will address the study of emotional events and their relation to daily routines.

A further step towards a better understanding of the image and sentiment analysis is needed, due to the subjectivity of what an image can invoke to different persons. To this aim, having annotations by different persons is critical to evaluate the inter- and intra-observer variability. Moreover, future experiments will address the generalization of the model over datasets collected by other wearable cameras, as well as recorded by different users.