Towards Generating Colour Terms for Referents in Photographs: Prefer the Expected or the Unexpected?

Sina Zarrieß and **David Schlangen**
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
first.last@uni-bielefeld.de

Abstract

Colour terms have been a prime phenomenon for studying language grounding, though previous work focussed mostly on descriptions of simple objects or colour swatches. This paper investigates whether colour terms can be learned from more realistic and potentially noisy visual inputs, using a corpus of referring expressions to objects represented as regions in real-world images. We obtain promising results from combining a classifier that grounds colour terms in visual input with a recalibration model that adjusts probability distributions over colour terms according to contextual and object-specific preferences.

1 Introduction

Pioneering work on natural language generation from perceptual inputs has developed approaches that learn to describe visual scenes from multimodal corpus data and model the connection between words and non-symbolic perceptual features (Roy, 2002; Roy and Reiter, 2005). In this paradigm, colour terms have received special attention. Intuitively, a model of perceptually grounded meaning should associate words for colour with particular points or regions in a colour space, e.g. (Mojsilovic, 2005). On the other hand, their visual association seems to vary with the linguistic context such as 'red' in the context of 'hair', 'car' or 'wine' (Roy and Reiter, 2005).

Recently, large-scale data sets of real-world images and image descriptions, e.g. (Young et al., 2014), or referring expressions (Kazemzadeh et al., 2014; Gkatzia et al., 2015) have become available and can now serve as a realistic test bed for models of language grounding. In this paper, we use the ReferIt corpus (Kazemzadeh et al., 2014) to assess the performance of classifiers that predict colour terms from low-level visual representations of their corresponding image regions.

A number of studies on colour naming have looked at experimental settings where speakers referred to simple objects or colour swatches instantiating a single value in a colour space. Even in these controlled settings, speakers use colour terms in flexible, context-dependent ways (Baumgaertner et al., 2012; Meo et al., 2014). Therefore, probabilistic models and classifiers, allowing for variable thresholds and boundaries between regions in a colour space, have been proposed to capture their grounded meaning (Roy, 2002; Steels and Belpaeme, 2005; Meo et al., 2014; Larsson, 2015).

Can we learn to predict colour terms for more complex and potentially noisy visual inputs? In contrast to simple colour swatches, real-world objects often have internal structure, their visual colour values are hardly ever uniform and the colour terms can refer to a specific segment of the referent (see image a) and b) in Figure 1). Moreover, the low-level visual representation of objects in real-world images can vary tremendously with illumination conditions, whereas human colour perception seems to be robust to illumination, which is known as the "colour constancy" problem (Brainard and Freeman, 1997). Research on colour perception suggests that speakers use "top-down" world knowledge about the prototypical colours of an object to *recalibrate* their per-

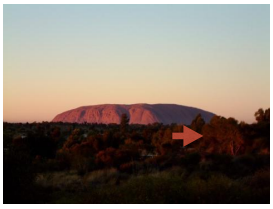
Proceedings of The 9th International Natural Language Generation conference, pages 246–255,
Edinburgh, UK, September 5-8 2016.

(a)"small red car on right"

(b"yellow building"

(c)"green"

(d)"first set of green on right"



(e)"red plants in the middle"

(f)"red rock bluff center"

Figure 1: Example images and REs from the ReferIt corpus

ception of an object to its expected colours (Mitterer and De Ruiter, 2008; Kubat et al., 2009). For instance, the use 'green' for the two, rather different hues in Figure 1 (c-d) might be attributed to the fact that both objects are plants and expected to be green.

However, recalibration to expected colours is not the only possible effect of context. Despite or because of special illumination conditions, the mountain in Figure 1 (f) and the plants in Figure 1 (e) are described as 'red', a rather atypical, unexpected colour that is, therefore, contextually salient and informative. This relates to research on referential over-specification showing that speakers are more likely to (redundantly) name a colour if it is atypical (Westerbeek et al., 2014; Tarenskeen et al., 2015).

In our corpus study, we find that these various contextual effects pose a considerable challenge for accurate colour term classification. We explore two ways to make perceptually grounded classifiers sensitive to context: grounded classifiers that are restricted to particular object types and "recalibration" classifiers that learn to adjust predictions by a general visual classifier to the preferences of an object and its context. Whereas object-specific colour classifiers perform poorly, we find that the latter recalibration approach yields promising results. This seems to be in line with a model by Gärdenfors (2004) that assumes context-independent colour prototypes which can be projected into the space of known colours for an object.

2 Grounding colour Terms: Visual Classifiers

In this Section, we present "visual classifiers" for colour terms that predict the colour term of an object given its low-level visual properties. We assess to what extent the visual classifiers can cope with the real-world challenges discussed above.

2.1 Corpus and Data Extraction

We train and evaluate on the ReferIt data set collected by Kazemzadeh et al. (2014). The basis of the corpus is a collection of "20,000 still natural images taken from locations around the world" (Grubinger et al., 2006), which was manually augmented by Escalante et al. (2010) with segmentation masks identifying objects in the images (see Figure 4). This dataset also provides manual annotations of region labels, with the labels being organised in an ontology (Escalante et al., 2010). Kazemzadeh et al. (2014) collected a large number of expressions referring to objects (for which segmentations exist) from these images (130k REs for 96k objects), using a game-based crowd-sourcing approach.

We extract all pairs of REs containing a colour word and their image region from the corpus. We consider REs with at least one of the 11 basic colour words *'blue', 'red', 'green', 'yellow', 'white', 'black', 'grey', 'pink', 'purple', 'orange', 'brown'*. We remove relational REs, containing one of the following prepositions: *'below', 'above', 'not', 'behind', 'under', 'underneath', 'right of', 'left of', 'ontop of', 'next to','middle of'* in order to filter instances where the colour term describes a landmark object. We split the remaining pairs into 11207 instances for training and 1328 for testing. Table 1 shows the frequencies of the colour adjectives in the training set.

2.2 Visual Input

Research in image processing has tried to define colour spaces and colour descriptors which are to

colour term	%	colour term	%
white	26.7	black	8.7
blue	20.5	brown	6.2
green	16.7	pink,orange	2.8
red	14.6	grey,purple	1.4
yellow	9.9		

Table 1: Distribution of colour words in training data

some extent invariant to illumination and closer to human perception, cf. (Manjunath et al., 2001; Van De Sande et al., 2010). As we are more interested in the linguistic aspects of the problem, we have focussed on the standard, available feature representations. We extracted RGB and HSV colour histograms for region segments with `opencv` (Bradski, 2000). As the region segments are sized differently, we normalised the histograms to represent relative instead of absolute frequencies.

Ideally, we would like to use a feature representation that could be generalised to other words contained in referring expressions. Therefore, we have extracted features that have been automatically learned with a high-performance convolutional neural network (Szegedy et al., 2015). We computed the smallest rectangular bounding box for our image regions, applied the ConvNet and extracted the final fully-connected layer before the classification layer. As bounding boxes are less precise than segmentation masks, it is expected that this representation will perform worse – but it gives us an interesting estimate as to how much the performance of our model degrades on visual input that is less tailored to colour terms. To summarise, we have extracted the following representations of our visual inputs:

- mean RGB values for region segment (3 features)
- RGB histograms with 512 bins (8 bins per channel) for region segment (512 features)
- HSV histograms with 512 bins (8 bins per channel) for region segment (512 features)
- ConvNet features for bounding box (1027 features)

2.3 Experimental Set-up

The task We define our classification problem as follows: input is a feature vector x, a visual representation of a referent in an image, and output is a label y, a colour term for the referent. For the sake of simplicity, we only consider training and testing instances that contain colour terms and do not model the decision whether a colour term should be generated at all. In standard NLG terminology, we are only interested in realisation, and not in content selection. A lot of research on REG has actually focussed on content selection, assuming perfect knowledge about appropriate colour terms for referents in a scene, cf. (Pechmann, 1989; Viethen and Dale, 2011; Viethen et al., 2012; Krahmer and Van Deemter, 2012; Koolen et al., 2013).

The classifiers We used a multilayer perceptron that learns a function from colour histograms (or ConvNet features) to colour terms, i.e. defining an input layer corresponding to the dimensions of the colour histogram and an output layer of 11 nodes. We did not extensively tune the hyper parameters for our different visual inputs, but tested some parameter settings of the perceptron trained on RGB histograms, singling out a development set of 500 instances from the training set described above. We report results for training on the entire training set with two hidden layers (240 nodes and 24 nodes), a drop out set to 0.2 and 25 epochs. When training on the mean RGB values as input, we use simple logistic regression as we only have 3 features.

We also tested a Knn (nearest neighbour) classifier which simply stores all instances of x in the training data, and during testing, retrieves the k instances that are most similar to the testing example based on some distance metric. We used the default implementation of Knn in `scikit-learn` (Pedregosa et al., 2011) which is based on Minkowski distance. Testing on the development set, we obtained best results with setting k to 10 and uniform weights (all neighbours of a testing instance treated equally).

Evaluation We report accuracy scores. When there are multiple colour terms for the same region, we use the top n predictions of the visual classifier.

2.4 Results

Table 2 reports the performance of the visual classifiers for the different visual inputs and the two classification methods. We see that Knn performs consistently worse than Perceptron. The ConvNet features perform dramatically worse than the colour histograms and do not even come close to a simple logistic regression trained on mean RGB values of

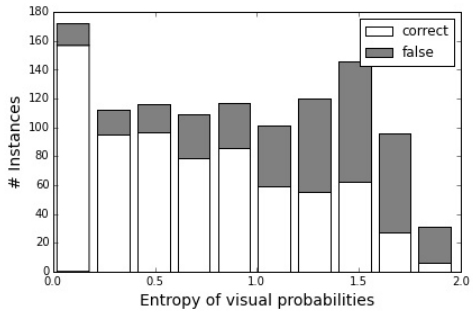


Figure 2: Proportion of correct vs. false predictions depending on the visual probability of the top-ranked colour term

the image regions. Surprisingly, we obtain better results with RGB histograms than with HSV.

	Perceptron	Knn
Mean RGB	57.29	55.65
RGB histogram (3d)	63.7	59.32
HSV histogram (3d)	62.84	55.73
ConvNet features	47.77	38.79

Table 2: Accuracies for general visual colour classifiers

Figure 2(a) shows the accuracy of the visual classifier depending on the (binned) entropy of the predicted probability distribution over colour terms. The accuracy (shown as the proportion of white and grey parts of a bar) is systematically higher in cases when the entropy is low, i.e. when the top colour has a clearly higher probability than the remaining colour candidates. This pattern suggests that the predicted probability distributions reflect the confidence of the visual classifier somewhat reliably. We consider this as evidence that the visual classifier learns to identify the prototypical instances of colour terms, whereas other, more ambiguous hues are associated with distributions of higher entropy.

2.5 Lexical vs. visual colour probabilities

Additionally, we assess the visual classifiers for different types of objects, based on the label annotations included in the corpus. We average the predicted visual probabilities for colour over all instances of an object label and compute the lexical probabilities of a colour term conditioned on the object label. These lexical probabilities tell us how often a colour co-occurs with a particular object label. Figure 3 shows the lexical and predicted visual probabilities (striped bars) for the labels 'flower', 'horse', 'hill', and 'car', illustrating some object-specific variation. For instance, flowers occur with many different colours, except "black", "brown" and "green". Horses, on the other hand, only occur with "white", "brown" and "black".

Depending on the object, the visual probabilities come more or less close to the lexical probabilities. The classifier predicts that flowers are more likely to be "green" than "blue", which reflects that flowers are likely to have certain green parts. The lexical probabilities, however, show a clear preference for "blue" over "green" since speaker mostly describe the salient, non-green parts of flowers. A more drastic case is "horse" where "brown" is frequent, but the classifier seems to systematically mis-represent this colour, predicting much more black horses than expected. For "hill", speakers almost exclusively use the colour "green" whereas the visual classifier predicts a flatter distribution among "blue", "green" and "white". As hills are often located in the background of images, the high probability for 'blue' certainly reflects a systematic, contextual illumination problem (see Figure 1(d) for a 'blueish' mountain).

Generally, the lexical colour probabilities in Figure 3 clearly show object-specific tendencies. In the following, we investigate how we can leverage that knowledge to adjust colour probabilities predicted on visual input to lexical preferences.

3 Object-specific Visual Classifiers

A simple way to make visual classifiers aware of object-specific colour preferences is to train separate classifiers for particular object types. This may not be a theoretically pleasing model for the meaning of colour terms, but in the following, we test whether this model improves the empirical performance for of colour term classification.

3.1 Object Types and Classes

Obviously, an object-specific model of colour terms crucially depends on the types of objects that we assume. How fine-grained does our object classification need to be? Intuitively, there are clear expectations about prototypical colours of certain objects (e.g. bananas vs. carrots), whereas other objects are more neutral (e.g. buildings, cars).

Fortunately, the ReferIt corpus comes with de-

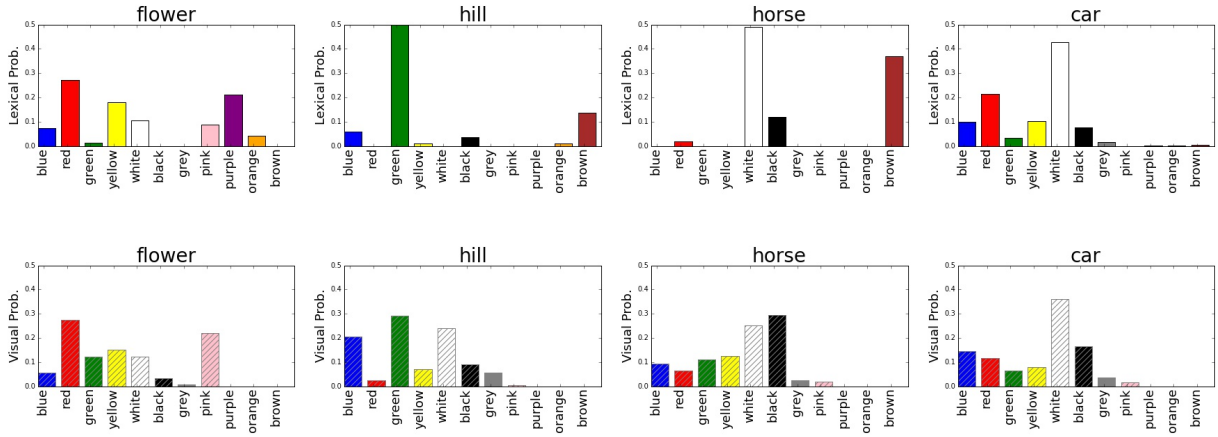


Figure 3: Lexical probabilities for colour terms conditioned on different types of objects (top row) and average visual probabilities predicted by the classifier trained on RGB histograms (bottom row)

tailed label annotations of the image regions (e.g. several types humans like 'child-boy', 'child-girl', 'face-of-person'). These object types are organised in an ontology, such that we can map relatively specific object type labels (e.g. 'car') to their general class (e.g. 'vehicle').[1] Table 3 shows the most frequent type labels and their classes in our training data. One very frequent type actually encodes colour information ('sky-blue' as opposed to 'sky-white' and 'sky-night' – leaves of the class 'sky').

object labels (classes)	# instances	top colour
man (humans)	1244	blue (22%)
woman (humans)	869	red (21%)
sky-blue (sky)	503	blue (98%)
group-of-persons (humans)	425	red (22%)
wall (construction)	421	white (42%)
car (vehicle)	418	white (42%)

Table 3: Most frequent object labels, their classes and most frequently mentioned colour (in the training instances of the visual classifier for colour)

3.2 Experimental Set-up

The Classifiers We use the same training data as in our previous experiment (Section 2.3). But now, we separate the training instances according to their labels (Section 3.1) and train several visual colour classifiers, i.e. one multi-class multi-layer perceptron per object label. In order to assess the impact of the underlying object classification, we used labels corresponding to (i) to the annotated, specific object types, (ii) the more general object classes. In each case, we only trained visual classifiers for labels with more than 50 instances in the training data. This leaves us with 52 visual classifiers for object types, and 33 visual classifiers for object classes.

Evaluation During testing, we assume that the object labels are known and we retrieve the corresponding visual classifiers. For objects with unknown labels (not contained in the training set) or an infrequent label (with less than 50 instances in the training set) we use the general visual classifier from Section 2.3 (the perceptron trained on RGB histograms). In Table 4, we report the colour prediction accuracy on the overall test set and on the subset of testing instances where the object-specific classifiers predicted a different colour term than the general visual classifier. This way, we assess how often the object-specific classifiers actually 'recalibrate' the decision of the general classifier and whether this calibration leads to an improvement.

3.3 Results

Table 4 shows that the classifiers trained for object types ($visual_{object}$) revise the decisions of the general classifier ($visual_{general}$) relatively often (for 619 out of 1328 testing instances), but rarely make a prediction that is different from the general classifier *and* correct (19% of the cases). Thus, overall, they severely decrease the performance of the colour term prediction. Similarly, the visual classifiers for object classes lead to a considerable decrease in performance. Interestingly, the predictions

[1] We map all object types below the node 'humans to 'humans'. Other categories on the same level are too general, e.g. 'man-made objects', 'landscape-nature' – here,we use the immediate mother node of the object label in the ontology.

Classifiers	# recalibrated colour terms	Accuracy on recalibrated subset visual$_{general}$	visual$_{object}$	Overall Accuracy visual$_{general}$	visual$_{general/object}$
Object types	619	57.9	19.	63.7	45.19
Object classes	357	72.54	8	63.7	45.58

Table 4: colour term prediction for general (visual$_{general}$) and object-specific (visual$_{object}$) visual classifiers, accuracies reported on the recalibrated subset where predictions differ between the general and the object-specific classifiers, and for the whole testset

of this model seem to often differ from the general visual classifier when the latter is relatively confident: the general visual accuracy on this subset is much higher (72%) than on the overall test set. This suggests that the object-specific visual classifiers do not learn prototypical meanings of colour terms and are much more sensitive to noise whereas the general colour classifier has an advantage rather than a disadvantage from seeing a lot of different instances of a particular colour.

4 Recalibrating Colour Terms

A model that generally adjusts its predictions to the expected colour terms for specific objects is clearly not successful. In this Section, we present an alternative approach that separates the grounding of colour terms on low-level visual from object-specific and contextual effects. Thus, instead of training object-specific colours directly on low-level visual inputs, we now learn to predict systematic adjustments or recalibration of the probability distributions that a robust general visual classifier produces.

4.1 Data preparation

In order to learn recalibrations of visual probability distributions over colour terms, we need training instances annotated with "realistic" output of the visual classifier (where the colour term with the highest probability does not necessarily correspond to the gold label). Therefore, we split our training data into 10 folds and apply 10-fold cross-validation (or so-called "jackknifing") on the training data, i.e. we have 10 folds that we annotate with a respective visual classifier trained on the remaining 9 folds.

4.2 Context-based Recalibration

So far, we have looked at the prediction of colour terms as a purely local problem. However, we expect other objects surrounding the target referent to have an effect on the selected colour terms, especially in cases where the visual classifier is less confident.

For each target region, we extract all the remaining distractor regions from the same image and apply the visual classifier. We compute a context vector by averaging over these regions and use the mean probability of each colour term. Based on the contextual colour probabilities, we can learn a function that adjusts the local probabilities for colour terms given additional evidence from the context.

The Classifiers We train logistic regression models for each colour term, where e.g. objects described as 'blue' are positive instances and objects described with a different colour are negative instances for the blue classifier. Instead of low-level visual input (colour histograms) we use the distributions over colour terms predicted by the visual$_{general}$ classifier as features and train the context-based recalibration on 22 features (11 probabilities for the region and 11 probabilities for the context).

4.3 Object-specific Recalibration

We can also model recalibration separately for each type of object. For instance, a recalibration classifier for 'horse' could learn that many horses classified as 'black' are actually referred to as 'brown' (see Section 2.5). Thus, we want to test whether object-specific recalibration classifiers learn to recover from systematic errors made by the general visual classifier for certain types of objects.

Combining object-specific and context-based recalibration could help to distinguish colours that are unusual and salient from unexpected colours that are due to e.g. specific illumination conditions. For instance, this classifier could learn that a 'blueish' hill is very unlikely to be blue, if there are a lot of other blue objects in the image.

The Classifiers For each object label, we train 11 regressions that adjust the probabilities of a colour terms predicted by the general visual classifier and whose training samples are restricted to instances of that object. We compare a simple object-specific

Recalibration	# recalibrated colour terms	Accuracy on recalibrated subset visual$_{general}$	recalibrated	Overall Acurracy visual$_{general}$	recalibrated
Context	135	43.7	40	63.7	63.3
Object types	193	38.3	42	63.7	64.5
Object classes	185	36.75	46.48	63.7	65.1
Object classes + context	201	34.32	46.26	63.7	65.57

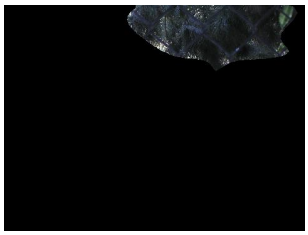
Table 5: Colour term prediction with context-based, object-specific and combined recalibration of the visual classifier, accuracies are reported on the recalibrated subset where predictions differ between the general visual classifiers and recalibrated colour terms, and for the whole testset

SUCCESSFUL OBJECT-SPECIFIC RECALIBRATIONS

(a) brown, visual: black

recalibrated: brown

INCORRECT OBJECT-SPECIFIC RECALIBRATIONS



(b) black, visual: black

recalibrated: green

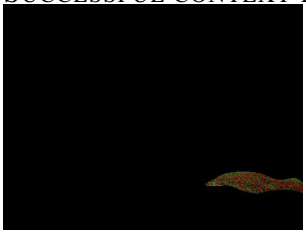
(c) green, visual: blue

recalibrated: green

(d) blue, visual: blue

recalibrated: white

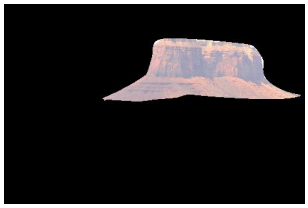
SUCCESSFUL CONTEXT-BASED RECALIBRATION



(e) red, visual: green

recalibrated: red

INCORRECT CONTEXT-BASED RECALIBRATION



(f) red, visual: pink

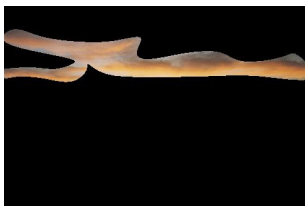
recalibrated: white

SUCCESSFUL COMBINED RECALIBRATION

(g): black, visual: red

recalibrated: black

INCORRECT COMBINED RECALIBRATION



(h) yellow, visual: yellow

recalibrated: white

Figure 4: Examples for successfully and mistakenly recalibrated colour term predictions, target regions on the left, full image on the right

recalibration that only takes the distribution over colour terms as input (11 features), and a combined recalibration based on a vector of 22 features (11 probabilities for the region and 11 probabilities for the context). Moreover, we train recalibration classifiers on object types (52×11 regressions) and object classes (33×11 regressions).

4.4 Results and Discussion

Evaluation We only recalibrate the visual probabilities for an object, if we have observed more than 50 training instances (same as in Section 3). For the remaining instances, we simply use the colour terms predicted by the general visual classifier. Thus, we will again be particularly interested in the subset of testing instances where the recalibration classifiers change the predictions of the visual classifier, which is the set of "recalibrated colour terms".

Table 5 shows the accuracies for the entire test set and the recalibrated subset. Except for the context-based recalibration which slightly degrades the accuracy compared to using only the visual probabilities (63.7%), the recalibration now improves the general visual classifier. The accuracies on the recalibrated subset reveal why recalibration is more successful than the object-specific visual classifiers discussed in Section 3: it is much more conservative in changing the predictions of the visual classifier. Moreover, the accuracy of the general visual classifier on the recalibrated test sets is substantially lower than on the overall test set. This shows that the recalibration classifiers learn to adjust those cases where the visual classifier is not very confident.

The accuracy of the visual classifier is not zero on the recalibrated subsets, meaning that some originally correct predictions are mistakenly recalibrated. Examples for correct and incorrect recalibration are shown in Figure 4, illustrating that the model has to strike a balance between expected and unexpected colour terms in context. There are several examples where the object-specific recalibration gives a higher probability to the more prototypical colour of the object (e.g. 'green' for trees and 'white' for houses in (a) and (c)), but this can lead to less salient, non-distinguishing or misleading colour terms being selected (Figure 4 (b,d)). The general context-based recalibration, on the other hand, often gives more weight to colours that are salient in the image (Figure 4(e)) , but sometimes calibrates the distribution in the wrong direction (Figure 4(f)). The combination of context-based and object-specific recalibration adjusts colour probabilities most reliably, and also seems to capture some cases of colour segments (Figure 4(g)). But there are still cases where the preference for expected or visually salient, unexpected colour is hard to predict, e.g. the "yellow cloud" in Figure 4(h).

These examples also suggest that an evaluation of the colour term prediction in terms of their interactive effectiveness might reveal different effects. The recalibration-based model lends itself for dynamic, interactive systems that adjust or correct their usage of colour terms based on interactive feedback.

Related Work Our notion of "recalibration" is related to a geometrical approach by (Gärdenfors, 2004) that separates colour naming conventions and prototypical, context-independent colour term meaning. Similarly, in distributional semantics, adjectives have been modeled as matrixes that map distributional vectors for nouns to composed vectors for adjective-noun pairs (Baroni and Zamparelli, 2010). Our recalibration classifiers can also be seen as first step towards modeling a compositional effect, but in our model, the noun (object label) adjusts the predictions of the adjective (colour). Finally, this works relates to research on vagueness of colour terms. But, instead of adjusting single thresholds between colour categories (Meo et al., 2014), the recalibration adjusts distributions over colour terms.

5 Conclusions

When speakers refer to an object in a scene, they often use colour terms to distinguish the target referent from its distractors. Accurate colour term prediction is thus an important step for a system that automatically generates referring expressions from visual representations of objects, cf. (Kazemzadeh et al., 2014; Gkatzia et al., 2015). This study has presented perceptually grounded classifiers for colour terms trained on instances of their corresponding referents in real-world images. We showed that this approach needs to balance various contextual effects (due to illumination, salience, world knowledge) and obtained promising results from a recalibration model that adjust predictions of a general visual classifier.

Acknowledgments

We acknowledge support by the Cluster of Excellence "Cognitive Interaction Technology" (CITEC; EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1193. Association for Computational Linguistics.

Bert Baumgaertner, Raquel Fernández, and Matthew Stone. 2012. Towards a flexible semantics: colour terms in collaborative reference tasks. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 80–84. Association for Computational Linguistics.

G. Bradski. 2000. OpenCV. *Dr. Dobb's Journal of Software Tools*.

David H Brainard and William T Freeman. 1997. Bayesian color constancy. *JOSA A*, 14(7):1393–1411.

Hugo Jair Escalante, Carlos a. Hernández, Jesus a. Gonzalez, a. López-López, Manuel Montes, Eduardo F. Morales, L. Enrique Sucar, Luis Villaseñor, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428.

Peter Gärdenfors. 2004. *Conceptual spaces: The geometry of thought*. MIT press.

Dimitra Gkatzia, Verena Rieser, Phil Bartie, and William Mackaness. 2015. From the virtual to the real world: Referring to objects in real-world spatial scenes. In *Proceedings of EMNLP 2015*. Association for Computational Linguistics.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, pages 13–23, Genoa, Italy.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

Ruud Koolen, Emiel Krahmer, and Marc Swerts. 2013. The impact of bottom-up and top-down saliency cues on reference production. In *Proceedings of the 35th annual meeting of the Cognitive Science Society (CogSci)*, pages 817–822.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Rony Kubat, Daniel Mirman, and Deb Roy. 2009. Semantic context effects on color categorization. In *Proceedings of the 31st Annual Cognitive Science Society Meeting*.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of logic and computation*, 25(2):335–369.

Bangalore S Manjunath, Jens-Rainer Ohm, Vinod V Vasudevan, and Akio Yamada. 2001. Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715.

Timothy Meo, Brian McMahan, and Matthew Stone. 2014. Generating and resolving vague color references. In *Proceedings of the 18th Workshop Semantics and Pragmatics of Dialogue (SemDial)*.

Holger Mitterer and Jan Peter De Ruiter. 2008. Recalibrating color categories using world knowledge. *Psychological Science*, 19(7):629–634.

A. Mojsilovic. 2005. A computational model for color naming and describing color composition of images. *IEEE Transactions on Image Processing*, 14(5):690–699, May.

Thomas Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27(1):89–110.

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

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(12):1 – 12. Connecting Language to the World.

Deb K Roy. 2002. Learning visually grounded words and syntax for a scene description task. *Computer Speech & Language*, 16(3):353–385.

Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language: a case study for colour. *Behavioral and Brain Sciences*, 28:469–489, 8.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR 2015*, Boston, MA, USA, June.

Sammie Tarenskeen, Mirjam Broersma, and Bart Geurts. 2015. Hand me the yellow stapler or Hand me the yellow dress: Colour overspecification depends on object category. page 140.

Koen EA Van De Sande, Theo Gevers, and Cees GM Snoek. 2010. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596.

Jette Viethen and Robert Dale. 2011. Gre3d7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of the UCNLG+ Eval: Language generation and evaluation workshop*, pages 12–22. Association for Computational Linguistics.

Jette Viethen, Martijn Goudbeek, and Emiel Krahmer. 2012. The impact of colour difference and colour codability on reference production. In *Proceedings of the 34th annual meeting of the Cognitive Science Society (CogSci 2012)*.

Hans Westerbeek, Ruud Koolen, and Alfons Maes. 2014. On the role of object knowledge in reference production: Effects of color typicality on content determination. In *CogSci 2014: Cognitive Science Meets Artificial Intelligence: Human and Artificial Agents in Interactive Contexts*, pages 1772–1777.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2(April):67–78.