

Exploring Human Judgement of Digital Imagery

Timo Volkmer

James A. Thom

S.M.M. Tahaghoghi

School of Computer Science and Information Technology

RMIT University

PO Box 2476V, Melbourne, Australia 3001

Email: {tvolkmer, jat, saied}@cs.rmit.edu.au

Abstract

Statistical learning methods are commonly applied in content-based video and image retrieval. Such methods require a large number of examples which are usually obtained through a manual annotation process, that is human raters review images and assign semantic concept labels. The human judgement, however, cannot be regarded as the ultimate truth because of its subjectiveness and the likelihood of human error. We can address these issues by using multiple judgements per example, but evaluating and resolving disagreement between raters is problematic. Moreover, the nature of rater disagreement and how to minimise it are not yet well explored. In this paper we present results of a user study that was specifically designed to investigate human judgement of digital imagery. We discuss the influence of factors such as size and type of semantic vocabulary on inter-rater agreement. We demonstrate the application of latent class analysis for combining multiple judgements. Known from applications in the medical and social sciences, this statistic allows robust, quantitative evaluation of multiple judgements per subject. We believe it is well suited for application during the evaluation and modelling phase in semantic image and video retrieval.

Keywords: image annotation, semantic annotation, latent class analysis, inter-rater agreement

1 Introduction

Supervised learning methods such as Bayesian networks or support vector machines are widely used for automatic image and video classification in multimedia information retrieval. Based on positive and negative training examples, such systems compute a binary model which can be used to classify unseen data automatically. The examples for training have to be annotated with semantic concept labels taken from a pre-defined vocabulary. The annotation task — performed manually by human reviewers — is time consuming and prone to errors. Moreover, the annotation is often based on only an individual opinion, that is only one judgement per image is obtained from a single rater. In this case, it is impossible to quantify the quality of the annotation unless another individual reviews all annotations.

The effectiveness of automatic classifiers depends on the underlying model, the quality of which in turn depends heavily on the training collection. This dependency is both qualitative and quantitative in

nature. Automatic classifiers need a large number of positive examples per concept but also require the annotation to be accurate and discriminative. Based on this need for large training collections, we want to minimise both annotation error and annotation time when generating annotated collections.

Given enough resources, we can generate multiple annotations per image by assigning several raters to the task. We expect an improvement of the annotation quality because we can model the view of several raters rather than only an individual opinion. However, we often observe significant disagreement between judgements. In the absence of a “gold standard” it is problematic to decide whether an image should be used as a positive or as a negative example. It is often impossible to classify an image into a category with 100% certainty as there is almost always room for ambiguity. Reverting to only using images that have unanimously been classified is a possible solution if the collection is large enough. But for rare concepts, or in smaller collections, this is likely to result in too few positive examples.

Some examples should perhaps neither be used as a negative nor as a positive example if there is much disagreement in their annotation. One might also decide to review the annotation in case the inter-rater agreement is below an acceptable level. Knowing that we almost always have to expect some disagreement when working with multiple judgements, we can quantify the annotation quality based on the level of inter-rater agreement. It is desirable to apply a robust statistic that supports both classifying annotated examples and estimating the rater agreement as a measure of annotation quality.

Driven by the above observations, we want to learn which factors influence the agreement between multiple human raters. We want to clearly classify individual images based on multiple judgements, and finally, we want to minimise the annotation time per image.

In this paper we present results of a user study addressing the reliability and efficiency issues of image annotation. As the problem of video retrieval is usually reduced to an image retrieval problem, we believe this work is applicable to both areas. We have explored different modes of annotation, and used different semantic concepts to study their impact on annotation quality. As a novelty in this field, we apply Latent Class Analysis (LCA) to compute concept prevalence and classification error, and compare it to the inter-rater agreement using intraclass correlation.

After presenting related research in the following section, we describe our experiment in Section 3, followed by an analysis of the results in Section 4. We conclude the paper with a discussion of the lessons learnt and an overview of our future work in Section 5.



Figure 1: Screenshot of the annotation system that we used. In multiple-concept mode all five concepts are activated, as shown here, while only one concept is activated for selection in single-concept mode.

2 Related Work

In prior work with others (Volkmer et al. 2005) we investigated some of the issues related to manual image and video annotation. Within the TRECVID¹ 2005 Annotation Forum, 61,904 key frames were collaboratively annotated by more than 102 reviewers. We used this opportunity to analyse the annotation including additional information in regards to timing and preferences of the input device. This gave us valuable clues, but it was difficult to reliably quantify concept frequencies and inter-rater agreement because the data was collected in an uncontrolled environment. For example, we were not able to determine whether annotations under one user account were indeed done by a single user. Moreover, not all images were reviewed by the same number of annotators. We hypothesised that the agreement between raters decreases if the judges try to annotate too many concepts simultaneously. Based on the data available, we were not able to confirm this. Many questions remained unanswered which led us to conduct a new experiment that is the subject of this paper.

Given the recent trend to share information on the Internet in blogs and online photo albums, we can find many examples in which large image collections are annotated. Most of them, however, do not use a controlled concept lexicon, and instead work with free-text annotations. Perhaps the most prominent example of such an approach is the Yahoo! photo sharing portal Flickr.² Each image can be assigned one or more tags describing some semantics of the image content. This annotation is highly subjective which makes a Flickr image collection less suitable for learning a specific concept. For example, the search term “Airplane” retrieves many images of people sitting in passenger aircraft or aerial views out of airplanes. These noisy results seem to occur because annotations are based on subjective cogitations that individuals connect with an image. The image con-

tent is not necessarily well reflected by the chosen terms. The ESP Game (von Ahn & Dabbish 2004) combines annotations by multiple users in a game-like application. Users annotate images that are found on the Internet; the system computes a confidence score based on how many different users agree on a certain description of an image. Users in turn collect points if they agree on an annotation. While this approach is very interesting, the annotation data is not available to us and we are unable to make a statement about its quality or usefulness for information retrieval purposes.

Other than the studies conducted in connection with the TRECVID Annotation Forums in 2003 and 2005 (Lin et al. 2003, Volkmer et al. 2005), we are currently unaware of any related studies for image annotation that use controlled concept vocabularies. However, analysis of survey reliability and response agreement is a well-studied area in the medical, behavioural, and social sciences. Latent Class Analysis (Lazarsfeld & Henry 1968) is an established method to evaluate survey data. Uebersax & Grove (1990) demonstrate its usefulness for analysing medical diagnostics agreement and extend this work beyond the medical sciences (Uebersax 1992, Uebersax & Grove 1993). The flexibility and robustness of latent class analysis has been demonstrated in other applications (Vermunt 1996, Vermunt & Magidson 2003, Vermunt 2003).

Deerwester et al. (1990) and Dumais (1995) use latent class analysis for an approach for indexing text documents and report improvements in retrieval performance over traditional term matching techniques. Hofmann (1999) extends this approach and proposes Probabilistic Latent Semantic Indexing (PLSI) for text document indexing. The idea underlying these indexing approaches is that the topic of a text document can be expressed as a latent variable that can only be indirectly obtained through modelling document terms as a response vector. In contrast to this work, we apply latent class analysis in evaluation and modelling of multiple judgements in annotated image collections.

¹<http://www-nlpir.nist.gov/projects/trecvid>

²<http://www.flickr.com>

Concept set a <i>vehicles</i>		Concept set b <i>settings/scenes/sites</i>	
a_1 Car	6%	b_1 Outdoor	33%
a_2 Truck	2%	b_2 Sky	13%
a_3 Bus	< 1%	b_3 Studio	11%
a_4 Airplane	1%	b_4 Building	12%
a_5 Boat/Ship	< 1%	b_5 Vegetation	10%

Table 1: The two sets of concepts, $a = \{a_1, \dots, a_5\}$ and $b = \{b_1, \dots, b_5\}$, that we have used in our experiment along with the expected prevalences based on prior experiments.

In the next section, we will describe our experiment before we discuss our findings in Section 4. We apply latent class analysis to compute concept prevalences and outline the possible usefulness of this statistic in the further process of semantic indexing of image collections.

3 Our Experiment

We designed the experiment to study human judgement of images in a controlled environment. We did not aim to generate an annotated collection of examples for training. Figure 1 shows a screenshot of the web-based application that we used. It was specifically designed for this task and allowed users to annotate one image at a time with either a single or multiple concepts.

To be able to relate our conclusions back to prior work (Volkmer et al. 2005), we selected images and semantic concepts for the annotation from the collection that was used in the TRECVID 2005 Annotation Forum. Based on the data collected in 2005, we were able to estimate the expected prevalences for the concepts that we use, as shown in Table 1. We describe the concept definitions and the image collection in detail below.

3.1 Methodology

We invited 20 users, 10 male and 10 female, to participate as annotators in our experiment. The participants were drawn from the general public and a pool of research students from the Science, Engineering, and Technology Portfolio at RMIT University. All users were familiar with the use of computers but mostly not experienced in the field of multimedia information retrieval. The experiment was conducted anonymously, that is each user was randomly assigned an anonymous user account so that responses could not be traced back to individual participants.

We allowed a brief training phase so that the participants could get used to the system before we started the experiment. All users were presented one image at a time and (depending on the annotation mode) either one concept or five concepts next to the image. The task asked to select all concepts that a user considered applicable to the image. The annotation system offers navigation buttons for users to go backwards and forward between images. However, we divided the image collection into several sets as described below. Free navigation forward and backwards was only allowed within one image set. We introduced this restriction to provide a minimum of guidance through the collection while allowing users to correct any errors they might make. Each user performed the experiment in two parts; one where they annotated multiple sets of images, each with a single concept that is varied between sets, and one where

Group	User	Part 1	Part 2
1	user1	a1A1, a2A2, a3A3, a4A4, a5A5	b1 b2 b3 b4 b5 B1
	user2	a1A2, a2A3, a3A4, a4A5, a5A1	b1 b2 b3 b4 b5 B2
	user3	a1A3, a2A4, a3A5, a4A1, a5A2	b1 b2 b3 b4 b5 B3
	user4	a1A4, a2A5, a3A1, a4A2, a5A3	b1 b2 b3 b4 b5 B4
	user5	a1A5, a2A1, a3A2, a4A3, a5A4	b1 b2 b3 b4 b5 B5
2	user6	b1 b2 b3 b4 b5 B1	a1A1, a2A2, a3A3, a4A4, a5A5
	user7	b1 b2 b3 b4 b5 B2	a1A2, a2A3, a3A4, a4A5, a5A1
	user8	b1 b2 b3 b4 b5 B3	a1A3, a2A4, a3A5, a4A1, a5A2
	user9	b1 b2 b3 b4 b5 B4	a1A4, a2A5, a3A1, a4A2, a5A3
	user10	b1 b2 b3 b4 b5 B5	a1A5, a2A1, a3A2, a4A3, a5A4
3	user11	a1 a2 a3 a4 a5 A1	b1B1, b2B2, b3B3, b4B4, b5B5
	user12	a1 a2 a3 a4 a5 A2	b1B2, b2B3, b3B4, b4B5, b5B1
	user13	a1 a2 a3 a4 a5 A3	b1B3, b2B4, b3B5, b4B1, b5B2
	user14	a1 a2 a3 a4 a5 A4	b1B4, b2B5, b3B1, b4B2, b5B3
	user15	a1 a2 a3 a4 a5 A5	b1B5, b2B1, b3B2, b4B3, b5B4
4	user16	b1B1, b2B2, b3B3, b4B4, b5B5	a1 a2 a3 a4 a5 A1
	user17	b1B2, b2B3, b3B4, b4B5, b5B1	a1 a2 a3 a4 a5 A2
	user18	b1B3, b2B4, b3B5, b4B1, b5B2	a1 a2 a3 a4 a5 A3
	user19	b1B4, b2B5, b3B1, b4B2, b5B3	a1 a2 a3 a4 a5 A4
	user20	b1B5, b2B1, b3B2, b4B3, b5B4	a1 a2 a3 a4 a5 A5

Figure 2: The specification of our experiment, each user annotated 360 images. For each image/concept pair, we obtain 4 ratings. Lowercase letters represent concepts, while uppercase letters each represent a set of 60 images.

they annotated one set of images with multiple concepts. For a given user, neither the images nor the concepts overlapped between parts.

The primary goal of our experiment was to study the effects on efficiency and inter-rater agreement of the following two factors:

- Concept vocabulary: Specific objects, such as *Car* or *Airplane*, that are less prevalent versus different shot-settings with higher prevalence, for example *Vegetation* or *Sky*.
- Annotation mode: Annotating one image in regards to a single concept at a time versus annotating one image with five concepts simultaneously.

During the experiment, we recorded the annotation generated by each user in a central database. Additionally, we measured the time that each user spent per image. After the experiment we asked users to complete a short survey with simple questions in regards to the annotation.

3.2 Specification

We randomly selected 600 images out of the TRECVID 2005 development collection (Over et al. 2005) for our experiment. This collection contains 61,904 key frames extracted from from U.S. American, Arabic, and Chinese television recordings. These recordings consist mostly of news programs interrupted by advertisement segments and some entertainment programs.

We divided the image collection into two equal parts, A and B , each comprising 300 images. Each set was then further divided into five subsets of 60 images each $A = \{A_1, \dots, A_5\}$, and $B = \{B_1, \dots, B_5\}$.

From the 44 concepts that Naphade et al. (2005) had prepositioned for the 2005 Annotation Forum for TRECVID, only 39 were finally used. We selected ten concepts out of these 39 and defined two

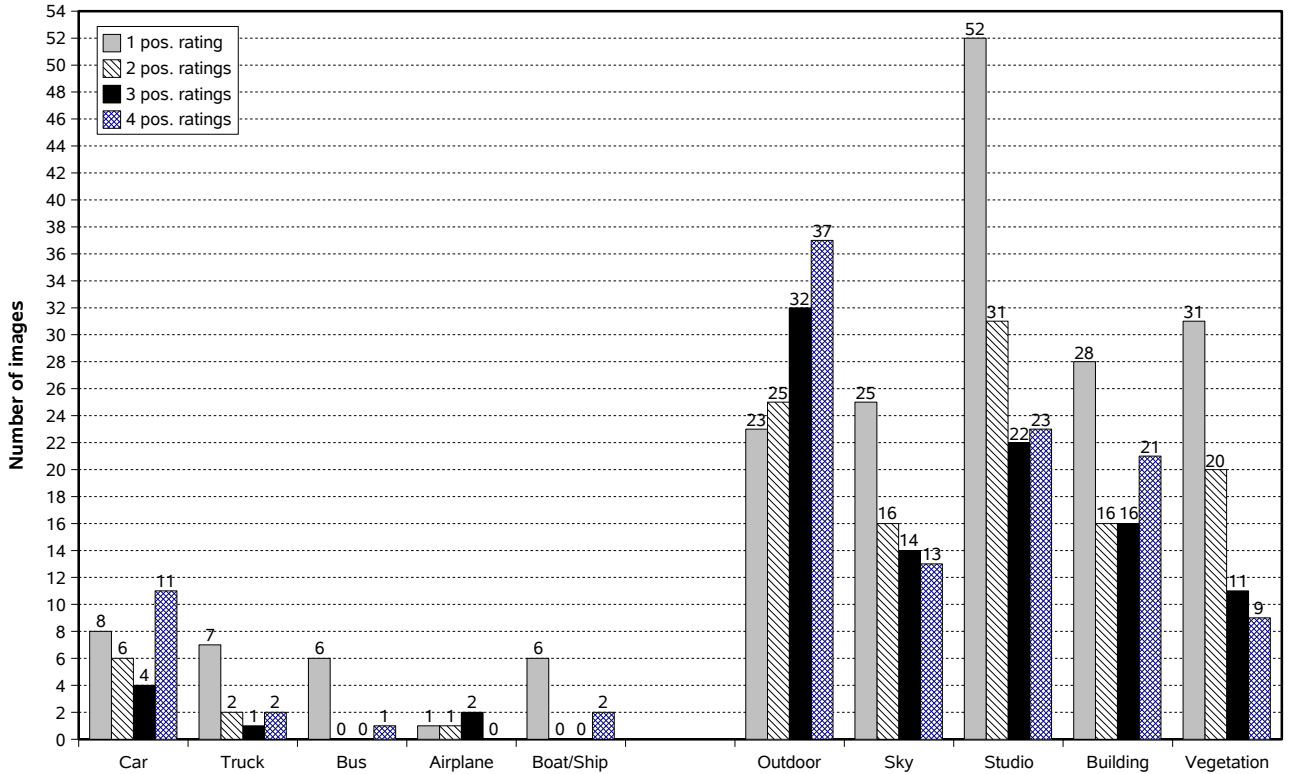


Figure 3: The raw positive ratings for each concept. Each bar represents the number of images that received 1, 2, 3, and 4 positive ratings for the respective concept. We observe substantial disagreement in the ratings.

sets a and b . These were selected such that one set $a = \{a_1, \dots, a_5\}$ represents well-known objects that appear rather rarely in the collection, while the second set $b = \{b_1, \dots, b_5\}$ consisted of settings and scenes with a significantly higher prevalence, as can be seen from Table 1.

Our experiment setup is illustrated in Figure 2. We grouped the 20 participants into four groups of five users each, and paid attention to achieving a uniform distribution of male and female participants among all groups. In Figure 2, a concept/image set combination such as a_1A_3 means that the user annotates the image set A_3 with concept a_1 (*Car*). Analogue to this, $b_1b_2b_3b_4b_5B_3$ means that the user annotates image set B_3 with all five concepts of concept group b . Each user would therefore annotate six subsets of images, that is 360 images, with all available concepts. As described above, the experiment was divided into two parts so that each user was to use both annotation modes. In single-concept mode, users were to annotate five subsets of 60 images each with one concept. The concept was varied for each subset. In multiple-concept mode, users were required to annotate one subset — 60 images — with five concepts, but this time with all five concepts simultaneously. To cancel out unwanted effects, we rotated the order in which users would see images among all users, and we rotated the order of annotation modes among user groups.

We believe this setup allows us to evaluate concept prevalences and inter-rater agreement based on the different annotation modes as well as in combination. Moreover, we can draw conclusions about which mode might be faster and we can make a statement about the impact of different concept types on the annotation.

In the next section, we discuss the results of our experiments after evaluating the data with latent class analysis. We conclude the paper summarising the lessons we have learnt in Section 5.

4 Evaluation of Experimental Results

The experiment setup results in four independent ratings per image, two for each single-concept and multiple-concept annotation mode. As not all users have annotated all 600 images with all available concepts, we can evaluate 300 distinct images for each concept based on four ratings.

Based on our prior experience, we anticipated the annotation speed to be faster if users have to annotate only one concept as opposed to five concepts at a time. We also expected the agreement between raters to be higher when annotating only a single concept. In addition to these effects, we expected users to have greater difficulty annotating the second set of concepts, set b , because these leave more room for ambiguity. We thus expected a higher disagreement rate for these concepts.

When evaluating the annotation, we are primarily interested in the positive ratings, that is the images that can be used as positive examples for training. All other images are usually considered negative examples. In our case, each image can obtain a maximum of four positive ratings. An optimal annotation would consist only of images with either zero or four positive ratings.

We counted the raw judgements, grouped by the number of positive ratings per image, and visualise these in Figure 3. Each concept is represented by four bars. Each bar shows the number of images that were judged as a positive example by 1, 2, 3, or 4 raters, respectively.

The graph in Figure 3 illustrates well the primary difficulty when using multiple ratings: there is significant disagreement in the number of positive judgements of most images. Consider concept *Airplane* in Figure 3 as an example. There is one image that one user has labelled as depicting an airplane. For another image, two raters have agreed that the image shows an airplane, however, two raters still disagreed

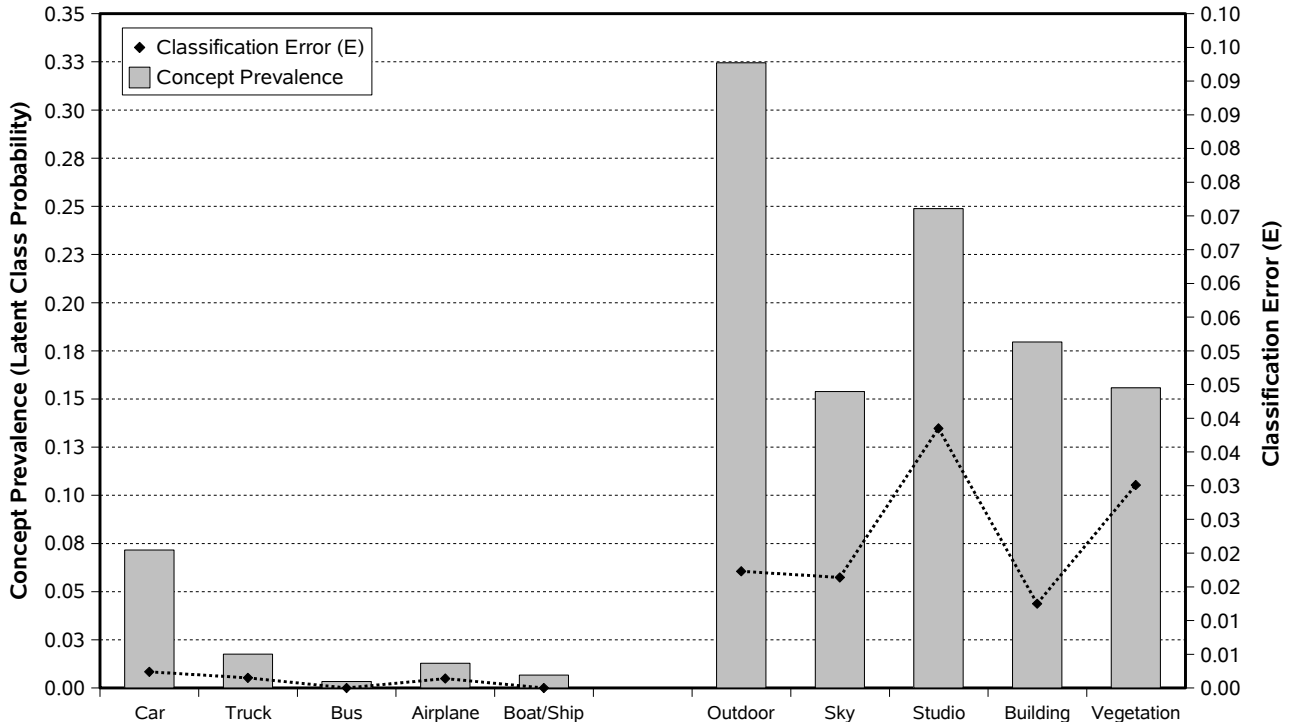


Figure 4: Concept prevalences, calculated using LCM, along with the estimated classification error E . We observe a generally higher error for the more frequent concepts of concept category b . E reaches 0 for concepts *Bus* and *Boat/Ship* because of the few positive examples for these.

with that. There are only two images that a majority of three raters labelled as depicting an airplane. And no single image was unanimously judged by all four raters as showing an airplane.

4.1 Latent Class Modelling

To solve this problem, we apply latent class analysis, or Latent Class Modelling (LCM) (Lazarsfeld & Henry 1968, Vermunt & Magidson 2003). This statistic allows us to combine multiple ratings per image and express these as a probability P that the respective image is indeed either positive or negative in regards to a particular semantic concept.

We adopt the basic idea of latent class modelling that the real classification of an image is contained in our annotation only as a latent variable X . This means that we cannot directly observe it, but we can derive it from the four observable ratings that we have for each image. We combine these ratings in the response vector \mathbf{Y} . According to the latent class modelling approach (Lazarsfeld & Henry 1968, Vermunt & Magidson 2003), the likelihood of obtaining a particular response pattern $P(\mathbf{Y} = \mathbf{y})$ for one item can be described as follows:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{x=0}^{C-1} P(X = x)P(\mathbf{Y} = \mathbf{y}|X = x) \quad (1)$$

Where C denotes the number of latent classes. We apply a restricted dichotomous model that assumes the two latent classes $x = 0$ (negative) and $x = 1$ (positive), expressed in the dichotomous latent variable X . $P(X = x)$ is the proportion of images belonging to the specific latent class x . This latent class model assumes mutual independence between all ratings for a latent class. The restriction that we apply is interchangeability of raters. This means that ratings of a particular image/concept pair can originate from

Concept set a		Concept set b	
<i>vehicles</i>		<i>settings/scenes/sites</i>	
a_1 Car	7.16%	b_1 Outdoor	32.45%
a_2 Truck	1.76%	b_2 Sky	15.39%
a_3 Bus	0.33%	b_3 Studio	24.88%
a_4 Airplane	1.28%	b_4 Building	17.96%
a_5 Boat/Ship	0.67%	b_5 Vegetation	15.58%

Table 2: Concept prevalences computed based on the new annotation obtained, using our two-class latent class model.

any of the 20 raters. The unrestricted model would consider all ratings per image to always be from the same raters, but this is not the case in our experiment.

To assign an individual image to a particular class, we apply the following Bayesian rule:

$$P(X = x|\mathbf{Y} = \mathbf{y}) = \frac{P(X = x)P(\mathbf{Y} = \mathbf{y}|X = x)}{P(\mathbf{Y} = \mathbf{y})} \quad (2)$$

We use modal classification, that is an image would be assigned to the class for which the highest probability $P(X = x | \mathbf{Y} = \mathbf{y})$ is computed.

Using the ℓEM (Vermunt 1997) program, we assessed our annotation with the two-class model described above. We computed the latent class membership probabilities for all concepts as values between 0 and 1. These can be interpreted as the likelihood of obtaining a positive example for the respective concept if one picked a random image out of our collection. The results are shown in Table 2, expressed as percentage values for each concept. For example, we observe 7.16% of the collaboratively judged images are rated as *Car*, that is approximately 21 images out of 300 depict one or more cars.

The results obtained through latent class modelling correlate well with the predicted concept preva-

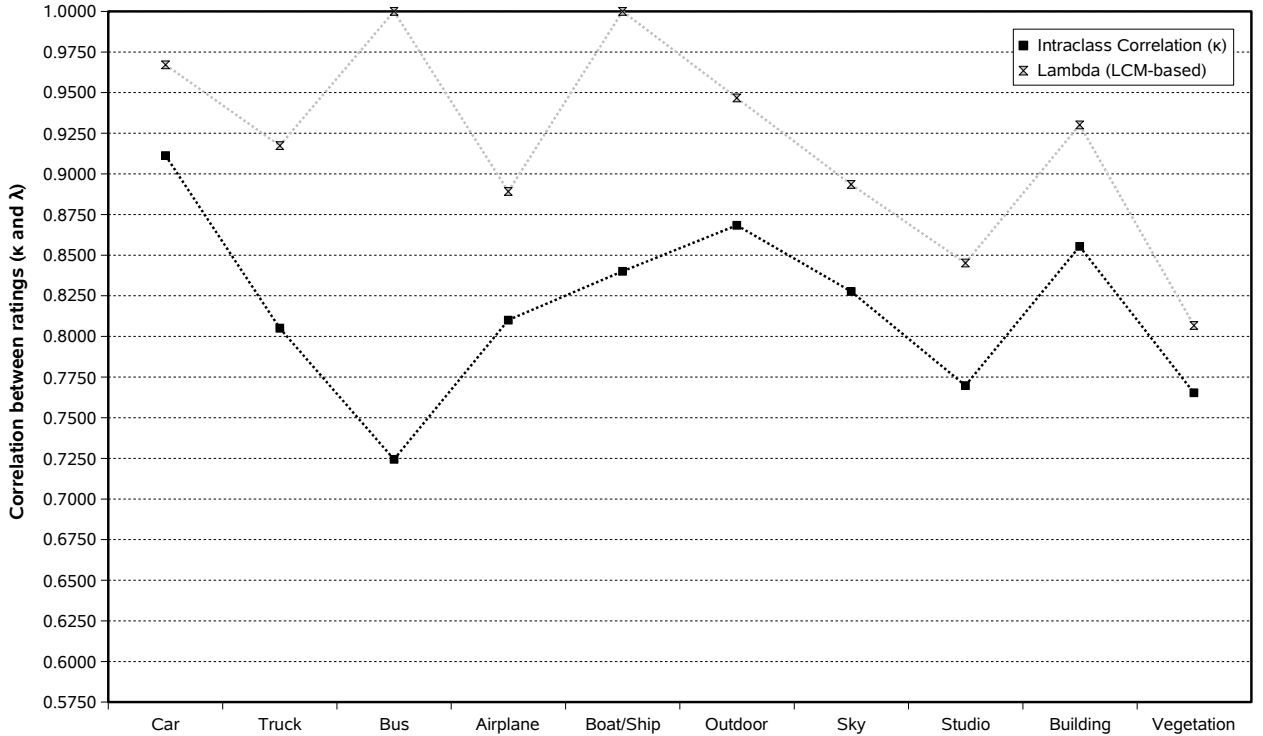


Figure 5: LCM classification performance λ compared to the intraclass correlation index κ . We observe a strong correlation between both: images can be better classified if there is less disagreement in how they have been rated. The concepts *Bus* and *Boat/Ship* are interesting outliers.

lences in Table 1. However, we believe the results computed with latent class modelling are more reliable, because our earlier results were only estimates based on a simple algorithm that counted the positive ratings for each image.

4.2 Estimating Classification Performance

An important measure for the classification performance of a model is the classification error E . It is the estimated proportion of false classifications based on all classifications for a particular concept. When using the modal classification rule as described in Equation (2), we estimate the classification error E for N images rated as follows:

$$E = \sum_{i=1}^I \frac{n_i}{N} \{1 - \max[P(X = x | \mathbf{Y} = y_i)]\} \quad (3)$$

Where I denotes the possible number of different response patterns and n_i the observed frequency for a particular response pattern.

Not surprisingly, the classification error has a direct relationship to the rater agreement because an image can be classified better if more raters agree. Figure 4 shows the computed concept prevalences graphed for each concept along with the classification error. In the case of the concepts *Bus* and *Boat/Ship*, the classification error is estimated to be 0. An explanation for this effect is found when examining the raw positive ratings shown in Figure 3. For the concepts *Bus* and *Boat/Ship*, we observe rather unambiguous response patterns that allow good classification. This is more likely to occur when only a few positive ratings are obtained. For *Boat/Ship* all four raters agree on “positive” in two cases, and in six cases three raters vote “negative” against a single rater. Almost identically, for the concept *Bus* all four raters agree on

“positive” in one case, while three raters vote “negative” against a single rater in six cases. The latent class model can resolve the ambiguity very reliably and the estimated error is $E = 0$, despite the fact that there is some disagreement.

However, we observe generally a higher classification error for the more frequent concepts, in particular those of concept set b . We have already confirmed this effect in prior work (Volkmer et al. 2005), that is we expect a higher disagreement in tandem with more prevalent concepts. Indeed, based on our results, we compute a correlation coefficient of $r = 0.81$ between concept prevalence and classification error — a strong correlation.

We can compare the LCM-based classification error E to the proportion of classification errors based on the unconditional latent class probabilities $P(X = x)$. This results in the classification performance measure λ , that is defined as follows:

$$\lambda = 1 - \frac{E}{\max[P(X = x)]} \quad (4)$$

As the measure λ is independent of class prevalence, we believe it can directly be used to assess annotation quality. However, λ is not a measure for inter-rater agreement. It estimates how well the example images could be classified based on the observed ratings; it should be interpreted similar to an R^2 measure (Vermunt & Magidson 2003). A value of $\lambda = 1.0$ indicates perfect classification.

To support our view, we compare λ to the intraclass correlation index (Fleiss 1981). This correlation index is a variant of the well-known Kappa statistic for inter-rater agreement by Cohen (Cohen 1960) and is therefore often referred to as Fleiss’ Kappa. In contrast to Cohen’s Kappa, however, the intraclass correlation index can be used to assess more than two judgements.

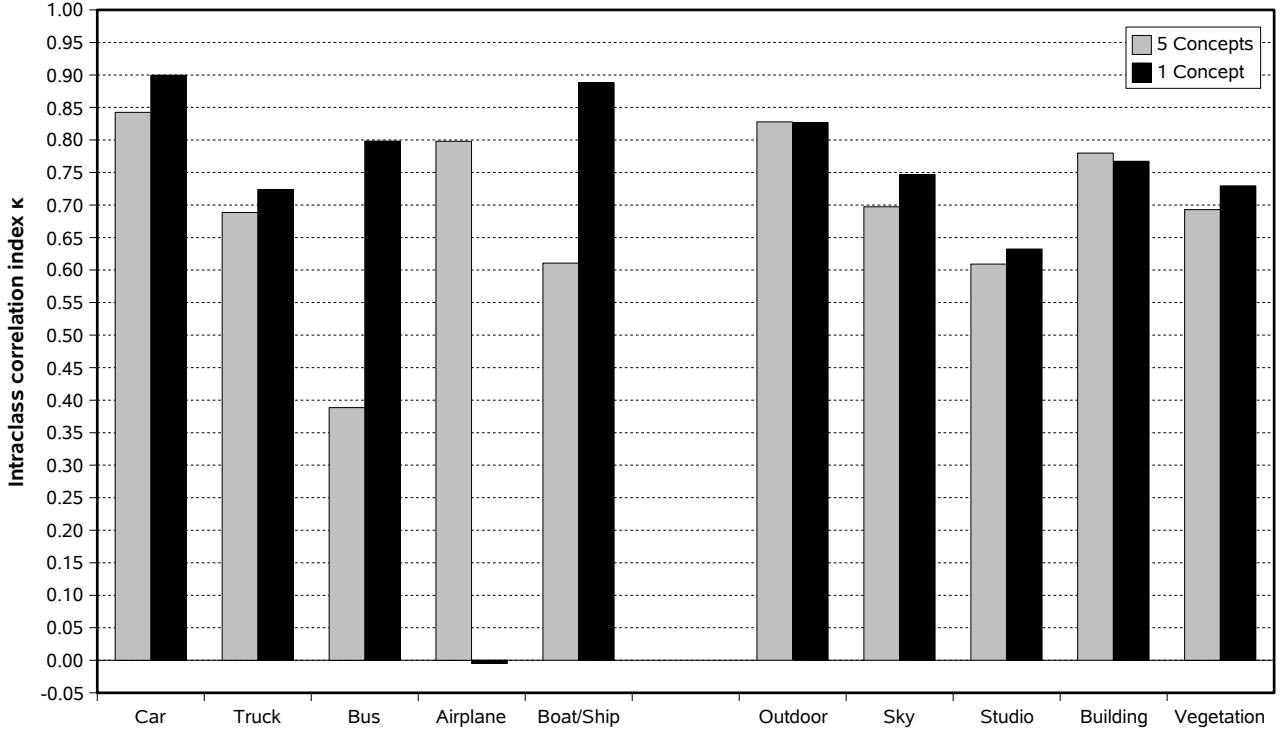


Figure 6: Intra-class correlation index κ for annotations done in single-concept mode compared to the intra-class correlation for annotations in multiple concept mode. We observe no statistically significant differences between both. For *Bus*, *Boat/Ship*, and *Airplane* in single-concept mode, the values for κ are most likely not reliable due to the very low prevalences of these concepts.

The intra-class correlation index is commonly denoted by κ and defined as follows:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \quad (5)$$

Where \bar{P} is:

$$\bar{P} = \frac{1}{N} \sum_{i=1}^N P_i \quad (6)$$

And \bar{P}_e is defined as:

$$\bar{P}_e = \sum_{j=1}^C p_j^2 \quad (7)$$

N is the number of samples, C is the number of latent classes. p_j is the proportion of all ratings to the j th of C classes. It is defined as follows:

$$p_j = \frac{1}{Nk} \sum_{i=1}^N n_{ij} \quad (8)$$

In Equation (8), n_{ij} is the number of judges who assigned the i th image to the j th class. The number of observed ratings per image is denoted by k , that is $k = 4$ in our experiment. The intra-class correlation index κ approaches 1.0 for perfect agreement. We compare κ and λ in Figure 5. The raw values for E , λ , and κ are shown in Table 3. Due to $E = 0$ for the concepts *Bus* and *Boat/Ship*, the LCM classification performance indicates perfect classification with $\lambda = 1.0$ in these cases. While there is some disagreement for these concepts, we believe that the low values of κ may not be reliable. As we pointed out above, the raw positive ratings do not imply as much ambiguity as κ suggests in these cases. For all other concepts, however, there is a strong correlation

between κ and λ . We compute the correlation factor for the eight most frequent concepts to be $r = 0.92$. Since we are not primarily interested in the level of rater agreement, but rather how well all images could be classified, we conclude that the LCM-based classification performance index λ is a good indicator for annotation quality when using multiple ratings.

Concept	E (LCM)	λ (LCM)	κ
Car	0.0024	0.9671	0.9113
Truck	0.0015	0.9174	0.8051
Bus	0.0000	1.0000	0.7244
Airplane	0.0014	0.8892	0.8101
Boat/Ship	0.0000	1.0000	0.8401
Outdoor	0.0173	0.9467	0.8684
Sky	0.0164	0.8935	0.8277
Studio	0.0385	0.8452	0.7698
Building	0.0125	0.9302	0.8554
Vegetation	0.0301	0.8068	0.7653

Table 3: Classification error E and classification performance λ based on latent class modelling compared to intra-class correlation index κ .

4.3 Inter-rater Agreement

We see from Figure 5 that the concepts *Bus*, *Studio* and *Vegetation* show the lowest agreement rates. However, the results for concept *Bus* in terms of κ may not be reliable because they are based on very few ratings. We therefore exclude it from the following analysis.

It appears that the concepts *Studio* and *Vegetation* were rather difficult to annotate for our participants. Indeed, 50% of all participants stated in the survey that they found it *difficult* or *very difficult* to annotate the concept *Studio*. On average, only 20% of all users perceived other concepts as difficult. Only 15% of all

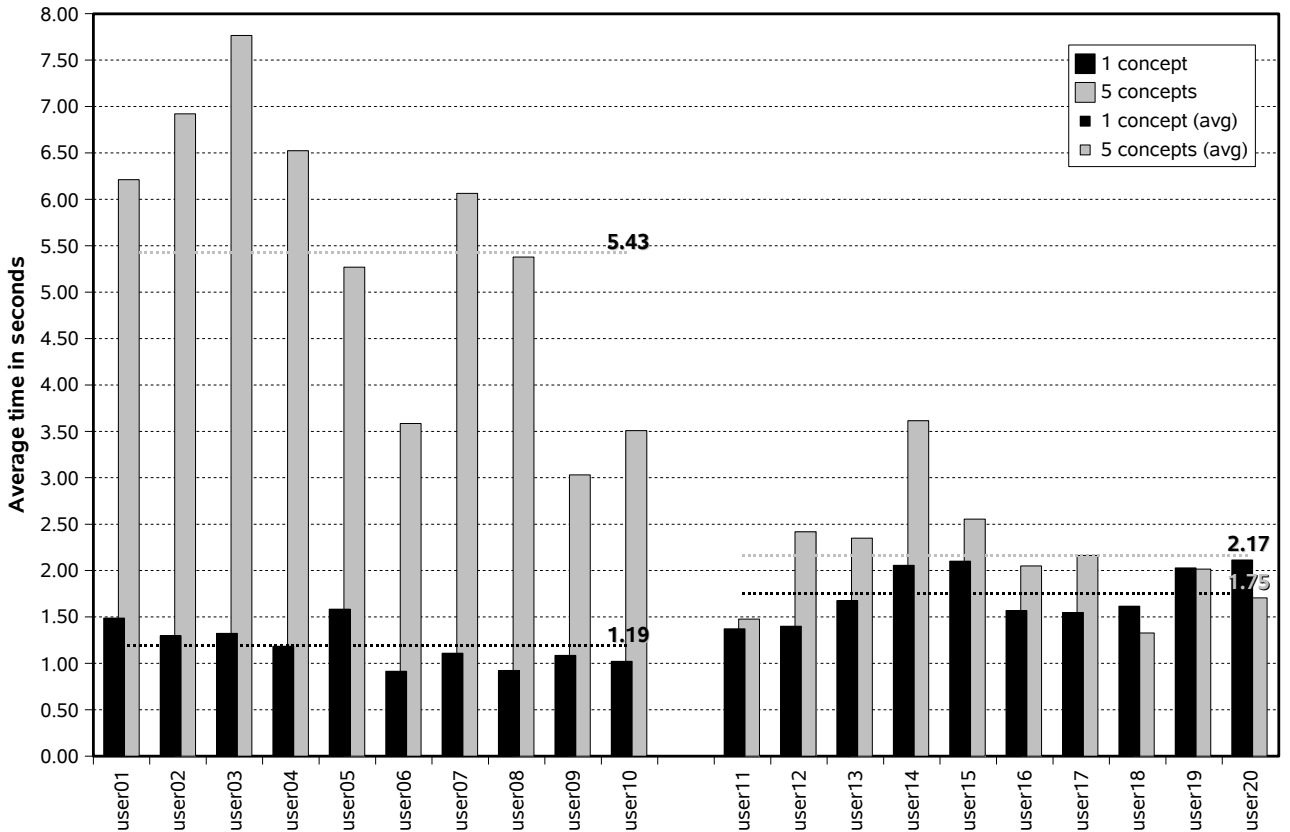


Figure 7: Annotation time per image by each user, naturally, only annotating one concept per image is faster. Interestingly, users 1 to 10 annotated their concepts in multiple-concept mode much slower than users 11 to 20. This is because users 1 to 10 annotated the more frequent concepts of group b in multiple-concept mode.

users rated *Vegetation* as difficult to annotate. However, this may only mean that users were not aware of its difficulty.

A potential weakness of our latent class modelling approach emerges when computing the model for very rare concepts, that is for sparse response tables (Vermunt & Magidson 2003). The LCM-based results for concept prevalence are highly likely to be reliable in such cases, but assessment of rater-agreement and annotation quality is difficult because the classification performance index λ may reach 1.0 despite notable disagreement. It is therefore important to understand λ as a purely statistical measure, that is as part of computing the model, no false classifications were estimated to have occurred. This does not guarantee, however, that no false classifications have happened in reality. Another problem of our model occurs when trying to apply it to fewer than three ratings per image. In this case, the system is overparameterised and we would need to introduce additional constraints. This does not seem possible in our case as it would mean, for example, that we have to assume one of our ratings to be absolutely reliable.

In the next step, we analyse the inter-rater agreement separated by annotation mode. For each image, we obtained four judgements, two that were done while the users annotated a single concept for each image, and two while multiple concepts were annotated. We use the intraclass correlation κ for this analysis. The results for κ in single-concept mode and multiple-concept mode are shown in Figure 6.

The computed value of $\kappa = -0.005$ for the concept *Airplane* in single-concept mode is not reliable because it is based on too few positive ratings. Only three positive ratings are observed, each of them in disagreement, which causes κ to drop to a value below zero. This may be interpreted as agreement be-

low chance, that is the agreement is lower than the agreement that can be expected by random assignment. Consistent with the problems that we previously observed when comparing κ and λ , we believe the κ measure is not necessarily reliable in cases with very few positive ratings. As can be seen from Figure 6, the differences for κ between single-concept mode and multiple-concept mode are most obvious for the three least prevalent concepts *Bus*, *Airplane*, and *Boat/Ship*. We will therefore treat these concepts as outliers in our analysis.

Regardless of some outliers, we observe differences between single-concept mode and multiple-concept mode. In most cases annotations done in single-concept mode have a higher intraclass correlation. While this seems to support our hypothesis that annotating several concepts at a time leads to higher disagreement, the differences we measure are not statistically significant.

Given that we are unable to reliably quantify the intraclass correlation for three of our concepts, and given that we cannot observe statistically significant differences for the other concepts, we cannot confirm this hypothesis. However, this conclusion is on the basis of using five concepts and we still believe that annotating too many concepts at a time may lead to higher disagreement. Further studies will be needed to determine a maximum number of simultaneous concepts that can be annotated while maintaining acceptable quality. Based on our current observations, we conclude that annotating five concepts at a time is well within the capacity of the average human annotator.

Similarly, we cannot confirm a dependency of the inter-rater agreement on the concept vocabulary based on our data. The average intraclass correlation for the concept sets a and b is nearly identical with

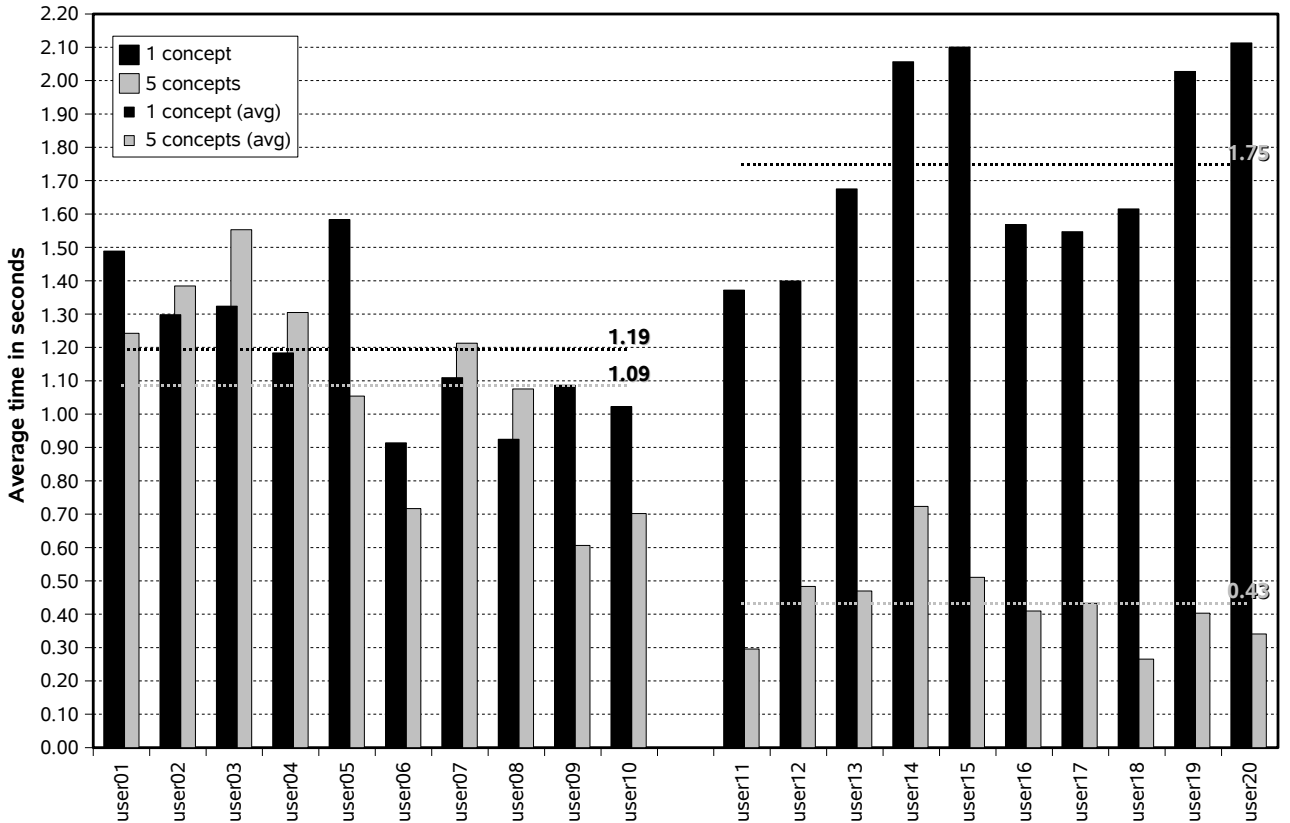


Figure 8: Annotation time calculated per image and per concept. We confirm that annotating the more frequently occurring concepts took on average longer than annotating the less frequent ones. Overall, annotating one image and one concept is done quickest when multiple concepts are annotated simultaneously.

$\kappa_a = 0.8173$ and $\kappa_b = 0.8182$. Unfortunately, the low frequencies of some concepts do not permit a more complete analysis. Further studies with a larger collection would be needed to draw more reliable conclusions.

4.4 Annotation Efficiency

We evaluated the timing data recorded during the annotation experiment. We measured the time that users spent while annotating each image. The average time that users spent per image is illustrated in Figure 7. We separated users 1 to 10 and users 11 to 20 into different groups and indicated their average times for both single and multiple-concept modes. From Figure 7 we can see that users 1 to 10 spent on average 5.43 seconds per image and five concepts while users 11 to 20 on average spent 2.17 seconds per image annotating five concepts. We believe the reasons for this effect is the fact that users 1 to 10 annotated the more prevalent concept group *b* in multiple concept mode. Users 11 to 20 annotated the concepts of group *a* simultaneously, these are significantly less frequent. For the same reason, users 1 to 10 were quicker in annotating in single-concept mode as they annotated the less prevalent concepts in this mode. Users 11 to 20 needed on average 1.75 seconds per image to annotate the more frequent concepts while users 1 to 10 needed on average 1.19 seconds for the less frequent concepts.

Our hypothesis is supported when evaluating the relative annotation times. We calculated the average annotation time needed per image and per concept in both modes and compared them in Figure 8. Again, we indicated the averages for users 1 to 10 and users 11 to 20 in both modes. We observe the shortest times for users 11 to 20 when they annotated the rare con-

cept set *a* in multiple-concept mode; only 0.48 seconds were on average needed per image and concept. Users 1 to 10 needed on average 1.09 seconds to annotate the more prevalent concepts of group *b* in combination. In conclusion, we can report that it is quicker to annotate in multiple-concept mode. Given that navigating between images in our annotation program causes an overhead, this is not surprising. Specialised implementations may address this by automatically guiding annotators to the next image after annotating one concept, but we do not believe that this will completely compensate the difference. Moreover, as we have confirmed that there is no significant increase in disagreement when annotating up to five concepts simultaneously, multiple concept annotation appears to be the preferable method to maximise efficiency. The average annotation time per concept and image among all participants in multiple-concept mode was 0.76 seconds. This is almost twice as fast as the 1.47 seconds that users on average needed to annotate in single-concept mode.

5 Conclusions and Future Work

We have conducted an experiment to study human judgement of digital imagery in regards to different annotation modes and pre-defined semantic concept categories. One goal of this study was to learn how to maximise efficiency while keeping disagreement between users to a minimum. Another goal was to explore the statistical combination of multiple user ratings for reliable image classification.

Overall, our results do not suggest any dependency of inter-rater agreement on different types of semantic concepts. However, individual concept specifications such as *Studio* or *Vegetation* need to be revisited as they seem to imply much ambiguity. Our initial hy-

pothesis, that annotating in multiple-concept mode leads to a decreased agreement between users, was not supported by our experiments. Annotating in multiple-concept mode may be the preferable strategy as long as the number of concepts does not exceed the capacity of the average annotator. Further experiments would be necessary to establish a threshold. In this light, and with regard to efficiency, it is preferable to annotate with multiple concepts simultaneously.

We have presented the application of latent class modelling for evaluating concept prevalences and classification performance. While the latter is not a measure of agreement, it strongly correlates with the inter-rater agreement and can serve as a quality index. We believe latent class analysis is a reliable statistic for this purpose and propose its application when training statistical learning algorithms. In particular, we propose the application of the LCM-based classification rules shown in Equations (1) and (2) to identify positive and negative examples. In the same way, LCM may be applied to combine results of multiple automatic classification algorithms during the retrieval phase. We believe that this can be helpful for implementing a combined retrieval approach. We will explore these applications of latent class modelling in greater depth in our future work.

As latent class models can be applied to varying numbers of raters (Uebersax & Grove 1990), we plan to revisit previously generated annotation data (Volkmer et al. 2005) for a new analysis.

We believe the outcome of this work forms the basis for better semantic indexing of image and video data.

Acknowledgements

We are very grateful to all 20 volunteers who have spent some of their valuable time to participate as annotators in our experiment.

References

- Cohen, J. (1960), 'A Coefficient of Agreement for Nominal Scales', *Educational and Psychological Measurement* **20**, 37–46.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. & Harshman, R. (1990), 'Indexing by Latent Semantic Analysis', *Journal of the American Society for Information Science* **41**(6), 391–407.
- Dumais, S. T. (1995), Latent Semantic Indexing (LSI): TREC-3 Report, in D. Harman, ed., 'NIST Special Publication 500-226: Proceedings of the Third Text REtrieval Conference (TREC 3)', Gaithersburg, MD, USA, pp. 219–230.
- Fleiss, J. L. (1981), *Statistical Methods for Rates and Proportions*, 2nd Ed., John Wiley & Sons, Inc., New York, NY, USA, chapter 13. The Measurement of Interrater Agreement, pp. 212–236.
- Hofmann, T. (1999), Probabilistic Latent Semantic Indexing, in D. Harman, ed., 'Proceedings of the ACM-SIGIR International Conference on Research and Development in Information Retrieval 1999', ACM Press, NY, USA, Berkeley, CA, USA, pp. 50–57.
- Lazarsfeld, P. F. & Henry, N. W. (1968), *Latent Structure Analysis*, Houghton Mifflin New York, New York, NY, USA.
- Lin, C.-Y., Tseng, B. L. & Smith, J. R. (2003), Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets, in E. M. Voorhees & L. P. Buckland, eds, 'TRECVID 2003 Workshop Notebook Papers', Gaithersburg, MD, USA. <http://www.alphaworks.ibm.com/tech/videoannex>.
- Naphade, M., Kennedy, L., Kender, J., Chang, S.-F., Smith, J. R., Over, P. & Hauptmann, A. (2005), A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005, Technical Report RC23612, IBM T.J. Watson Research Center, Hawthorne, NY, USA. [http://domino.watson.ibm.com/library/CyberDig.nsf/papers/A33ABDB65967B5%3B852570070056B36F/\\$File/rc23612.pdf](http://domino.watson.ibm.com/library/CyberDig.nsf/papers/A33ABDB65967B5%3B852570070056B36F/$File/rc23612.pdf).
- Over, P., Ianeva, T., Kraaij, W. & Smeaton, A. F. (2005), TRECVID-2005 – An Introduction, in P. Over & T. Ianeva, eds, 'TRECVID 2005 Workshop Notebook Papers', Gaithersburg, MD, USA. <http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tv5intro.pdf>.
- Uebersax, J. S. (1992), 'A Review of Modeling Approaches for the Analysis of Observer Agreement', *Investigative Radiology* **27**(9), 738–743.
- Uebersax, J. S. & Grove, W. M. (1990), 'Latent Class Analysis of Diagnostic Agreement', *Statistics in Medicine* **9**, 559–572.
- Uebersax, J. S. & Grove, W. M. (1993), 'A Latent Trait Finite Mixture Model for the Analysis of Rating Agreement', *Biometrics* **49**, 823–835.
- Vermunt, J. K. (1996), Log-linear Models for Event Histories, PhD thesis, Department of Methodology and Statistics, Tilburg University, The Netherlands.
- Vermunt, J. K. (1997), LEM: A General Program for the Analysis of Categorical Data, Technical report, Department of Methodology and Statistics, Tilburg University, The Netherlands.
- Vermunt, J. K. (2003), Applications of Latent Class Analysis in Social Science Research, in T. D. Nielsen & N. L. Zhang, eds, '7th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU) 2003', Vol. 2711 of *Lecture Notes in Computer Science*, Lecture Notes in Computer Science, Aalborg, Denmark, pp. 22–36.
- Vermunt, J. K. & Magidson, J. (2003), 'Latent Class Analysis', *The Sage Encyclopedia of Social Sciences Research Methods* **3**, 549–553.
- Volkmer, T., Smith, J. R., Natsev, A., Campbell, M. & Naphade, M. R. (2005), A Web-based System for Collaborative Annotation of Large Image and Video Collections, in 'Proceedings of the ACM International Conference on Multimedia 2005', Singapore, pp. 892–901.
- von Ahn, L. & Dabbish, L. (2004), Labeling Images with a Computer Game, in 'Conference on Human Factors in Computing Systems (CHI) 2004', Vienna, Austria, pp. 319–326. <http://www.espgame.org>.