Automatic grouping of semantic keywords to improve image rendering

Albrecht Lindner [†], *Nicolas Bonnier* [‡], *Mehmet Candemir* [†], *and Sabine Süsstrunk* [†]
[†] *School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*
[‡] *Océ Print Logic Technologies, Créteil, France*

Abstract

The ultimate goal of automatic image rendering is a system that gives at least as pleasing results as a human expert using an image manipulation program. In this article we demonstrate that the exploitation of semantic image keywords is a promising approach towards this ultimate goal. We develop a keyword classification scheme specifically for the purpose of automatic image rendering. Further on, we propose a method to automatically classify keywords into these classes. We discuss the results based on experiments with a database of 40'000 images, annotated on average by five keywords each.

Introduction

Enhancing digital images to make them visually more appealing is an important aspect in digital photography. Many software tools exist for this task, but due to the semantic gap – the fact that computers don't understand semantic context as well as human beings do – they do not work automatically but need human guidance. Let us consider an algorithm for enhancing the attractiveness of human faces by warping them to make the face appear more symmetric [9]. This is good in many cases, but wrong if a facial expression is desired that does not match common standards of beauty (e.g. frowning one's brow). Unlike a computer, a human being would recognize that the frowning look is essential and either leave it asymmetric or make it even more apparent.

In the context of this work, we define image rendering as either color rendering [1] or photo enhancement (e.g. adjustment of color, contrast, sharpness) that is either applied globally or locally to an image. We focus specifically on semantic based image rendering. Thus, either the whole image or different regions of an image are processed according to the image's or regions' specific content.

Content aware image processing is not a new topic. Cameras exploit user settings for internal processing of images if e.g. portrait mode is chosen or if the user defines the light source for white balancing. Technical metadata can also be used for indoor/outdoor classification [3]. Ciocca et al. propose a system that uses different classifiers and detectors to estimate the content of an image and base further processing on that information [4]. These examples show that technical metadata and automatic classifiers can add some semantic information, although it is very limited and on a much lower level in comparison to the semantic understanding of a human being.

A different and promising approach towards automatic image rendering is to gather and analyze semantic metadata that comes along with an image file (see Figure 1 for an example) and base further processing on the so gained information. Adding semantics has already proven to help other imaging related problems, such as object recognition [10, 13] or image retrieval [15]. The vocabulary is not controlled and users are free to enter anything that comes to mind when looking at the image. Thus, keywords can describe objects, colors, feelings and so forth. They are therefore a potentially valuable and reliable source for semantic information. A correct processing of this information has great potential to improve automatic image rendering.

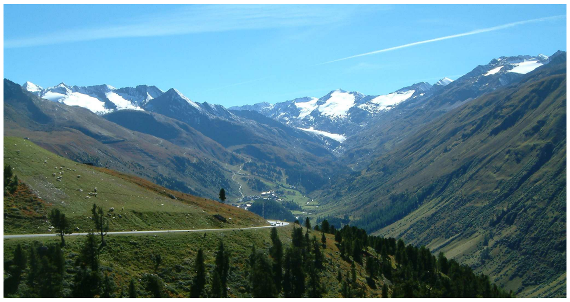


Figure 1. *Example image with annotated keywords: trees, green, mountains, snow, quiet, blue sky, road.*

A first step to handle the very diverse lexicographic input from keywords is to categorize them depending on the kind of semantic information they contain. Thus, the goal of this work is the organization of semantic metadata from keywords for the specific purpose of better automatic image rendering. This work is based on real world data from a large database of photographic images [14] and the proposed methods are inspired by and evaluated with it.

In this article we first discuss and propose an appropriate classification scheme for the given context. We give example images for the different classes and explain how they influence automatic image processing. Then we explain how we preprocess keywords with tools from natural language processing in order to simplify the classification task. We show how WordNet – a lexical database – can be used to efficiently classify keywords using our proposed classification scheme. We finish with an evaluation and critical discussion of the performance of the proposed classification system.

Images and keywords

The standard on photo metadata from the International Press Telecommunications Council (IPTC) [8] defines keywords in the context of photos as follows:

> Keywords to express the subject of the image. Key-

words may be free text and don't have to be taken from a controlled vocabulary.

Due to the broad definition, users can freely express their thoughts when looking at an image. The string entered by a user is stored in the keyword field of the IPTC header. Other sources for keywords related to an image can be found in other text fields of the image file's header, the filename, and the local surrounding of the image in a compound document.

In this project we use a database of 40'000 photographic images from 10'000 photographers. The images have been collected during The Flux project [14]. This project was realized in conjunction with the Musée de l'Elysée de Lausanne and the New York Photo Festival. The photographers were asked to upload and annotate their images and in return, their images were put on display in a photo exhibition of the participating museums. The images in the database have five keywords per image on average.

An investigation of the 300 most frequently used keywords in the database showed that there are only 2.3% adjectives and 0.5% verbs. The remaining 97.2% are nouns. Thus we chose to limit ourselves to nouns and assume that we dispose of a converter that gives the corresponding noun for a given non noun (e.g. [happy] → [happiness]) [1].

The keywords were preprocessed with three standard methods from computational linguistic:

- Compounds such as [stone age] were interpreted as a single expression.
- Stoplists were used to discard words that are due to the grammatical structure such as [for, the, and].
- Stemming was used to reduce inflected words to their stem, e.g. [trees] → [tree].

We used functions and word lists provided with a linguistic Perl package [7].

Keyword classes

Different keywords may influence varying parameters of automatic image rendering. Grouping keywords into distinct classes depending on their meaning for automatic rendering is thus an essential first step. It is important to define the context for which classes are meant to be used since this strongly influences how the classification scheme will be built. The IPTC definitions are meant to be used in news and press context. This does not match with our context – image rendering – and thus we propose a classification scheme for this very purpose. An optimal classification scheme assigns to every possible element at the input a single class. However, this is not always possible (or necessary) as explained in the following discussion.

We start building the classification scheme with the purpose of improving automatic image rendering. One of the first clear distinctions between rendering algorithms is whether they are applied globally or locally on an image. Hence, our keyword classification scheme has to account for this. Figure 2 illustrates a class diagram, where we separate the classes according to global or local characteristics.

We subsume all keywords that indicate a localizable object within an image in a first class denoted {object}[1]. A special subclass of this is formed by keywords that describe persons. It is justifiable to define a new class {person} since there are some specific characteristics related to persons. First, skin color is a memory color and thus needs special attention. Second, persons in images (e.g. friends or relatives) have a special relevance to the viewer.

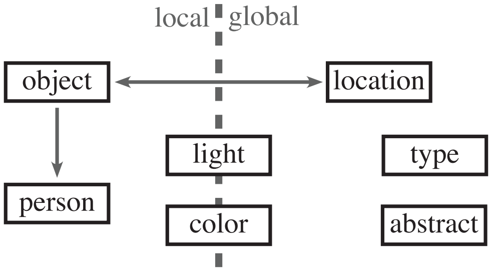


Figure 2. *Illustration of classification scheme.*

Closely related to the class {object} is the class {location}. This can be explained by means of a simple example keyword. Photos annotated with the keyword [airplane] can either show an airplane or can be taken in an airplane, but without showing it. In our database are 42 images that show an airplane or parts of it, 28 images are taken from an airplane and show anything except an airplane, and 13 images are taken from an airplane and show at least a part of the airplane (in most cases the wing). Other examples of that kind are [car, train, beach, house, mountain].

Thus we define the first three classes as follows:

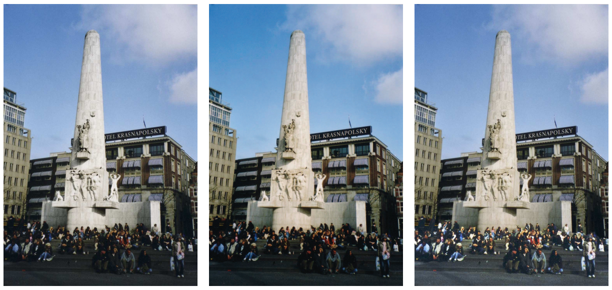
• {object}, natural or man-made, e.g. [tree, car]: Keywords from this class can be located in an image with object detection algorithms [6]. As previously discussed, attention has to be payed to objects that could also be used as a location. Once an annotated object is localized in an image, it can be highlighted with specific rendering since it is a priori an important part of the image. Highlighting could be achieved by increasing luminance or contrast as shown in Figure 3.

• {person}, e.g. [woman, Thomas]: This class is a subset of the class {object} since a person is also a localizable object. In addition to the rendering options discussed for the class {object}, special attention has to be payed to skin color and red eyes. An example image with a group of persons is depicted in Figure 3 on the right.

• {location}, e.g. [Paris, England]: Keywords from this class can not always be used for a semantic analysis of an image. There is no different rendering intent for an image if it is taken in a forest and is annotated with either [England] or [France]. However, in some cases the location can be exploited. This is the case when the location is well known or very typical for a specific look. Let us consider two images with the title [Night in Las Vegas] and [Night in Atacama desert]. The first image has very likely colored light sources whereas the second does not. Such an example is given in Figure 4.

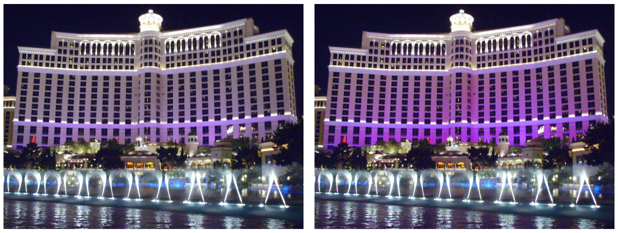
The next classes that we introduce are {color} and {light}. It is important to note that there are also keywords related to time such as [night, noon]. The time is important for rendering in the sense that it gives hints on the lighting conditions. For example,

[1] To improve readability, in the following we put all keywords in squared brackets: [keyword] and all keyword classes in curly brackets: {keyword class}.



[column] original [friends]

Figure 3. Example image showing different rendering for classes {object} on the left and {person} on the right. In both cases the region containing the important object has been lightened.



original [Las Vegas, night, magenta, color]

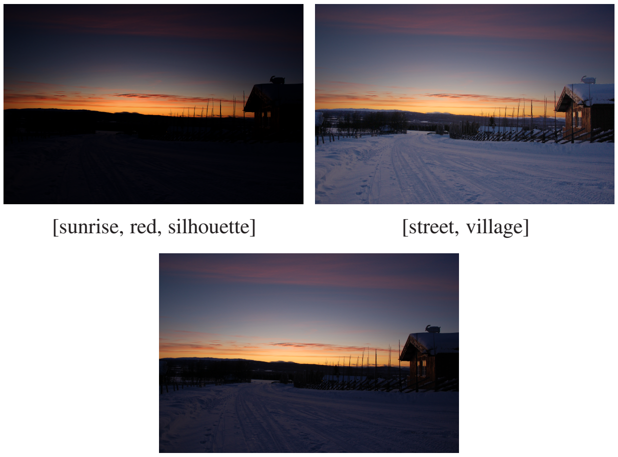
Figure 4. Example image showing rendering for classes {location, time, color} on the right.

keyword [night] means that the illuminating light source is artificial or faint (moon- or starlight). We therefore add keywords related to time also to the class {light}.

- {color}, e.g. [colorful, red, black and white]: Keywords from this class can be of local or global nature. It is local if there is an object (or region) in the image with a predominant color such as [red skirt]. In this case the additional color information of an object can be used to localize it. Global examples of this class are [sepia, black and white]. The image rendering can be optimized to amplify the dominance of the concerned color as in Figure 4.
- {light}, e.g. [sun, night, sunset]: Keywords from this class can also be local or global. Local, if the source is visible in the image (e.g. [sun]) and global, if the source is not visible but the scene has been illuminated by it (e.g. [moonlight]). Information about the light source under which an image has been taken is crucial for finding the white point. A priori knowledge about the light source's color temperature can be used for automatic white balancing. Keywords of this class are also linked to the class {color}. For example, an image with keyword [sunrise] provides the information that red is probably a predominant color in the image. An example is given in Figure 5.

Finally we define two truly global classes denoted {type} and {abstract}.

- {type}, e.g. [portrait, macro, silhouette]: Keywords from this class describe the type of image and they give strong indications what to expect in the image. The keyword [portrait] indicates that the image shows a frontal and centric view of a person's face, which facilitates its detection. Another example is given in Figure 6 where the keywords are [flower, depth of field]. This indicates that the flower is the main object of the image and that the rest should be blurred out. Yet another example is the keyword [silhouette] in Figure 5.

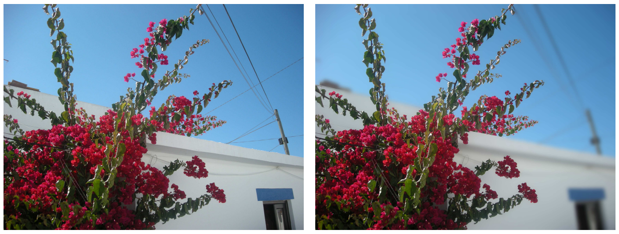


[sunrise, red, silhouette] [street, village]

original

Figure 5. Example image showing different rendering for classes {object} on the left and {light, color, type} on the right.

- {abstract}, e.g. [fun, wedding, hate]: This class gives an indication of the atmosphere of the image. This can be expressed by emotions such as [love, dolefulness] or indirectly by events such as [wedding, war]. Happy events could need a rendering that produces crisp and light colors. On the other hand, sad events could be more acceptable with more gentle colors. We point out that we did not define a class {event} since the event itself is not relevant for adaptive image rendering.



original [flower, depth of field]

Figure 6. Example image showing processing for classes {object, type} on the right.

Discussion of keyword classes

The classification scheme that we proposed in this section is still coarse and can of course be further refined for a more specific application. For example, it could make sense to split up the class {object} into classes {natural object} and {man-made object}, or to split up classes {light} and {color} into subclasses global and local. But we believe that the scheme in Figure 2 is sufficient for a discussion of keyword classification in the context of image rendering.

The ambiguity between the classes {object} and {location} is challenging. It is hard to define a rule that predicts how likely it is for a keyword to belong to the one or the other class. Of course, all objects that people can not – or normally do not – enter are purely of class {object} (e.g. [apple, closet]). Further on, an investigation of the database showed that annotations with names of countries, regions, and cities (e.g. [Italy, Colorado, Paris]) belong in almost all cases to the class {location}. Yet, for some keywords further information is necessary to find the right class. In this case the context has to be taken into account; e.g. other keywords or the image content may help to estimate the correct class for each particular case [12, 2].

We discuss our classification scheme on the basis of the 50 most used keywords in our database. According to our scheme, 45 can be assigned to the different classes as follows (the keywords within each class are listed with decreasing number of occurrence):

- {object}: water, flower, tree, landscape, architecture, sky, snow, cloud, animal, building, bird, shadow, boat, cat
- {person}: child, woman, girl, people, man
- {location}: city, street, Italy, New York, Paris, Switzerland
- {light}: light, sun, sunset, night, winter
- {color}: black and white, color, blue, red, white, green, black
- {type}: portrait, self portrait, macro
- {abstract}: nature, travel, love, beauty, heaven

There are five keywords that could not be classified. An investigation of the database showed that people use them as members of different classes. These keywords are [beach, lake, mountain, sea, reflection]. The first four can occur as members of {location} or {object}. The keyword [reflection] is ambiguous. It could e.g. be a reflection of an object or just a specular reflection on the surface of an object. In these cases the right class has to be determined from the context, such as other keywords or visual image content.

Automatic keyword classification

For automatic processing of an image it is necessary to have a machine-driven classification of keywords. The challenge is that people do not limit themselves to a fixed set of keywords when annotating images (see definition in the corresponding IPTC standard [8]). Hence it is necessary to have a classification algorithm that is flexible enough to handle this versatile input. For this purpose, we propose to use a lexical database that defines semantic relations between words. An example extract of a lexical database is illustrated in Figure 7. It defines hypernym and hyponym relations. In Figure 7 *plant* is a hypernym (generalization) of *tree*. The terms *oak* and *beech* are hyponyms (specializations) of *tree*.

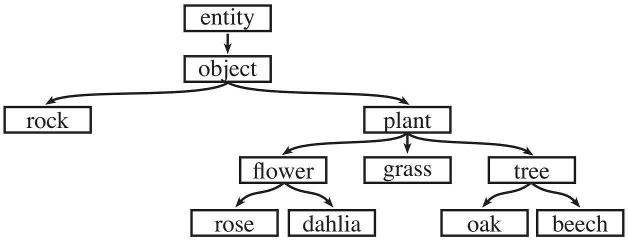


Figure 7. *Part of a lexical database in a tree structure.*

One well known lexical database for the English language is WordNet by Miller and Charles [11], and it is widely used in language processing. The object detection community also discovered it as a handy tool and started using it in the form of ImageNet[5]. We decided to use Wordnet release 3.0 due to its availability and its wide acceptance in the community.

In WordNet, each node is called a synset. It is important to point out that a node is not equal to a word but to a sense. For example, the word *tree* is represented in three different synsets: 1) the plant 2) the diagram 3) an actor called *Sir Herbert Tree*. WordNet orders the synsets with decreasing probability of appearance, which can be retraced in the before mentioned example. There exist word sense disambiguation techniques that deal with the problem of finding the right sense of a word in a given context [12]. For the moment we do not use such a system and thus take the first sense in WordNet, which gives the highest probability to guess the right sense.

In the next subsection we propose an approach to automatically classify a keyword. The evaluation is done with the keywords from The Flux database [14]. All keywords from the 40'000 images have been extracted and sorted with descending frequency. Keywords that appear in three or less images have been suppressed, which leaves 3527 different keywords.

Classification via hypernyms

WordNet's tree structure is already a grouping of senses. In the example of Figure 7 it becomes evident that [rose], [dahlia], [oak] and [beech] have the hypernym synset *plant* in common and are thus member of class {plant}. This concept can be extended to the case where a class is represented by several hypernym synsets instead of a single one. A keyword [keyword] is then member of class {class} if one of its representing synsets is a hypernym of [keyword].

Classification via hypernyms is very easy to implement and does not need parameter tuning. However, the hypernyms have to be carefully chosen. In this section we discuss our choice for every class of Figure 2. We start with the easier classes and end with the more difficult ones.

Color

We chose for the class {color} the synset with the sense *color, colour, coloring, colouring*. Based on this definition the following keywords have been identified as class members: [color, blue, red, green, black, pink, yellow, sepia, gray, purple, brown, sky blue, beige, scarlet, amber, coral, fawn, ebony, magenta, crimson]. The only missing keyword is [white], which is due to the fact that WordNet's first sense of this word is *caucasian*. The images in the database showed that the keyword [white] in a large number of cases is related the color and not the person. Hence this keyword is an exception from our assumption that WordNet's first sense guesses the right sense. Wrongly classified is the keyword [fawn], which people use for a young deer instead of the color.

Location

The class {location} has the particularity that some of the keywords are ambiguous and could also be part of the class {object}. This ambiguity can not be modeled with WordNet since this is simply not what it was designed for. Hence, we limit the classification to those keywords that are clear members of that class: names of countries, regions, cities and so forth. We thus chose the two synsets *district, territory, territorial dominion, do-*

minion and *land, dry land, earth, ground, solid ground, terra firma*. In total 292 keywords have been classified as member of this class and the 30 most used are: [city, Italy, New York City, Paris, Switzerland, Japan, London, India, Lausanne, Usa, France, China, Geneva, Australia, Africa, Spain, Brazil, California, Mexico, Canada, Germany, Brooklyn, island, Argentina, Thailand, Rome, Venice, Texas, Barcelona, Manhattan].

Light

The classification results for the class {light} are best with the synset *electromagnetic radiation, electromagnetic wave, nonparticulate radiation*. The following keywords have been classified as a member: [light, sunlight, sunshine, glowing, moonlight, ray, sunbeam, candlelight]. In our definition of this class we argued that keywords relating to time are also member of this class since they may indicate the lighting conditions. The two synsets we chose for this are: *hour, time of day* | *morning, morn, morning time, forenoon*. This adds the following keywords to this class: [sunset, sunrise, morning, dawn, dusk, twilight, sundown, aurora, rush hour, sun set, midnight, daybreak].

Abstract

The class {abstract} is for keywords related to emotions and events that typically indicate emotions. Thus we use the following three synsets: *feeling* | *condition, status* | *social event*. The 30 most frequently used class members are then: [love, cold, documentary, atmosphere, concert, sleep, film, poverty, shoes, joy, happiness, pollution, silence, campaign, ruin, emotion, clear, heart, hope, fear, wet, mystery, wedding, race, melancholy, celebration, sadness, passion, soil, curiosity]. This class has a very broad definition and thus needs several hypernym synsets. Most of the keywords belong to this class with a few exceptions: [shoes, soil]. Obviously, stemming did not work for the keyword [shoes], the reason is that WordNet has one sense for this word and thus does not reduce the stem to the more obvious word [shoe]. The keyword [soil] is not necessarily wrong. WordNet's first sense is *dirt* and most of the images from the database with this keyword effectively express the abstract concept *dirtiness*.

Dealing with the classes {light} and {abstract} revealed an issue with WordNet. If a class is very broad, it's hypernym synset needs to be far up in the tree hierarchy in order to account for the class' diversity. The drawback of this is that more and more wrong detections are made since the hypernym becomes too general and it subsumes too many words. The alternative approach is to choose several hypernym synsets that are lower in the tree hierarchy. The lower one goes the more hypernyms are necessary to cover the whole width of that class.

Object

The class {object} is challenging due to the very same reason. We chose four hypernym synsets which are: *object, physical object* | *flora* | *matter* | *animal*. Additionally, we excluded all keywords that have already been classified as a member of one of the other classes. With this approach the 30 most frequently used keywords are: [water, flower, tree, landscape, architecture, beach, sun, sky, street, mountain, animal, building, bird, boat, cat, heaven, colors, dog, plant, house, window, bridge, garden, wall, orange, park, rock, ice, wood, forest]. In this list there are two wrong detections: [colors, orange]. The first one is interpreted in the sense of *flag* and the second one in the sense of *fruit*. However, an investigation of the database showed that people mostly use those keywords in the sense of color.

Person

For the class {person} we have chosen the synset with the sense *person, individual, someone, somebody, mortal, soul*. With this synset, 254 keywords of the database are classified as a member of that class and the 30 most used are: [child, woman, girl, white, man, boy, baby, dali, kid, skipper, friend, photographer, mother, tourist, modern, homeless, natural, youth, daughter, lady, architect, gull, crane, artist, musician, father, sweet, contemporary, tiger, dancer]. The keyword [white] is wrongly classified as previously discussed. An investigation of [photographer] showed that 83% of the images actually do show a person, though not always with a camera. For [tourist] it is at least two thirds. The keywords [modern, natural, contemporary] are not used as nouns. But since we use WordNet only for nouns it returns as first sense a person with that characteristic. This issue can be resolved by incorporating non nouns in the classification and estimating the probability that a given word is rather used as a noun or something else. As previously discussed this concernes only 2.8% non nouns in the database. Further on, there are issues with proper names that can be summarized with the three keywords [Dali, Obama, Crane]. The first one is a person (artist) but people mean his paintings, whereas the second stands for the person itself. The third keyword [crane] is more often used as a lifting device than as the writer's name *Stephen Crane*.

Type

We were unable to find a good set of hypernym synsets for the class {type}. In order to avoid too many false positives the hypernyms had to be chosen that low in the tree, that it ended up being a list of all keywords of that class. Our list of manually chosen keywords is: [portrait, self portrait, macro, photography, photo, blur, contrast, nude, long exposure, still life, silhouette, street photography, close up, exposure, photograph, digital image, skyline, digital art, drawing, digital photography, fisheye, infrared, symmetry, portraiture, panoramic, blurred].

Unclassified keywords

The classification described above does not have overlapping classes within our database. However, there are keywords that remain unclassified. The first 50 in decreasing order of frequency are: [black and white, people, night, lake, snow, winter, cloud, reflection, travel, shadow, sea, environment, life, abstract, urban, music, face, summer, wed, river, family, spring, old, rain, ocean, eyes, fun, holiday, church, autumn, sport, view, dark, fire, fog, culture, movement, storm, evening, foot, beautiful, peace, colorful, smile, tourism, construction, solitude, freedom, leman, market].

In total, 52% of the keywords of the database have been classified. If the keywords are weighted by their occurrence in the database, the classification rate is 63%. This means that more frequently used keywords are more often classified and less frequently used keywords remain more often unclassified.

Conclusions and Future Work

Even though many images are annotated with keywords, today's image processing rarely exploits them. In this article we investigated the possibility to use semantic keywords for automatic image rendering and enhancement. Keywords contain a lot of information about an image e.g. its content, color and light characteristics, mood and what the photographer intended to express with it. All this is potentially rich information worth investigating if it can be incorporated into an automatic image processing workflow.

We showed that keywords can be grouped into classes depending on how they can influence automatic image rendering. Based on this discussion we proposed a classification scheme specifically designed for this task. The scheme's basic division is global versus local keyword classes. This is due to the fact that the same division can be done for image rendering algorithms. We defined seven classes and illustrated with several example images how they could be used in automatic image rendering.

Finally, we proposed an automatic keyword classification method. The main challenge to correctly classify keywords is that it is not controlled vocabulary since people are free to enter any text that comes to mind when looking at an image. This is necessary in order to give them enough freedom to describe their thoughts and feelings, but makes it also very difficult for automatic processing. To account for that, we proposed to use WordNet for the classification since it covers a large vocabulary of the English language and provides valuable semantic relationships between words.

The classification algorithm was tested on The Flux database consisting of 40'000 manually annotated images from 10'000 photographers. The classification performed well on a majority of the cases and is a promising approach to handle such diverse lexicographic input.

The main drawback of the current implementation is the rather high rate of unclassified keywords of 37% (occurrence weighted average). This rate can be lowered by adding more hypernyms to the definition of a class. The class consists then of more subtrees and becomes larger. However, these hypernyms have to be carefully chosen in order to avoid increasing false detection rate. There is a trade-off between increasing the classification rate while avoiding misclassification. Because we understand semantic image rendering to be an optional post-processing step, we prefer a higher rate of unclassified keywords over misclassification.

In the future, we expect a significant gain in precision by relating keywords. So far, our proposed algorithm processes every keyword by itself. The annotation [red skirt] is split up and then classified as {color} and {object}. The semantic information that both keywords belong to the same thing in the image is lost. More sophisticated natural language processing techniques are necessary to extract such semantic information.

We will also investigate how to link the proposed classification scheme to established image rendering techniques. For this purpose it will be necessary to study how rendering techniques influence the semantic image content described by keywords of different classes.

Author Biography

Albrecht Lindner is a PhD student in the School of Computer and Communication Sciences at EPFL (Lausanne, Switzerland). He works on improving print quality by means of high-level image processing. His research is sponsored by Océ Print Logic Technologies. In 2008, Albrecht obtained his MS degrees in Electrical Engineering, concentration Signal and Image Processing from the University of Stuttgart (Germany) and Télécom Paristech (France).

References

[1] ISO 22028-1. *Photography and graphic technology – Extended colour encodings for digital image storage, manipulation and interchange – Part 1: Architecture and requirements*, 2004.

[2] Kobus Barnard and Matthew Johnson. Word sense disambiguation with pictures. Technical Report TR-04-12, University of Arizona Computer Science Department, 2004.

[3] Matthew Boutell and Jiebo Luo. Bayesian fusion of camera metadata cues in semantic scene classification. In *CVPR*, volume 2, pages 623–630, Washington DC, June 2004.

[4] Gianluigi Ciocca, Claudio Cusano, Francesca Gasparini, and Raimondo Schettini. Content aware image enhancement. In *Artificial Intelligence and Human-Oriented Computing*, volume 4733/2007, pages 686–697, Rome, September 2007.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, pages 248–255, Miami, June 2009.

[6] http://pascallin.ecs.soton.ac.uk/challenges/VOC/. Pascal visual object classes challenge, visited June 2010.

[7] http://www.cpan.org. Wordnet similarity 2.05.

[8] International Press Telecommunications Council. *Standard on Photo Metadata*. www.iptc.org, July 2009.

[9] Tommer Leyvand, Daniel Cohen-Or, Gideon Dror, and Dani Lischinski. Data-driven enhancement of facial attractiveness. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2008)*, 27(3), August 2008.

[10] Marcin Marszalek and Cordelia Schmid. Semantic hierarchies for visual object recognition. In *CVPR*, pages 1–7, Minneapolis, July 2007. IEEE Computer Society.

[11] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[12] Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):1–69, February 2009.

[13] Devi Parikh and Tsuhan Chen. Hierarchical semantics of objects (hsos). In *ICCV*, 2007.

[14] Daniel Tamburrino, Patrick Schönmann, Patrick Vandewalle, and Sabine Süsstrunk. The Flux: Creating a Large Annotated Image Database. In *IS&T/SPIE Electronic Imaging: Image Quality and System Performance V*, volume 6808, San Jose, CA, USA, January 2008.

[15] Changbo Yang, Ming Dong, and Farshad Fotouhi. Learning the semanticsc in image retrieval - a natural language processing approach. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 137–143, Detroit, June 2004.