How 'How' Reflects What's What: Content-based Exploitation of How Users Frame Social Images

Michael Riegler[2], Martha Larson[3], Mathias Lux[1], Christoph Kofler[3]
[1]Institute for Information Technology, University of Klagenfurt, Austria
[2]Media Performance Group, Simula Research Laboratory AS, Norway
[3]Multimedia Computing Group, Delft University of Technology, Netherlands
michael@simula.no, {m.a.larson, c.kofler}@tudelft.nl, mlux@itec.aau.at

ABSTRACT

In this paper, we introduce the concept of *intentional framing*, defined as the sum of the choices that a photographer makes on *how* to portray the subject matter of an image. We carry out analysis experiments that demonstrate the existence of a correspondence between image similarity that is calculated automatically on the basis of global feature representations, and image similarity that is perceived by humans at the level of intentional frames. Intentional framing has profound implications: The existence of a fundamental image-interpretation principle that explains the importance of global representations in capturing human-perceived image semantics reaches beyond currently dominant assumptions in multimedia research. The ability of fast global-feature approaches to compete with more 'sophisticated' approaches, which are computationally more complex, is demonstrated using a simple search method (SimSea) to classify a large (2M) collection of social images by tag class. In short, intentional framing provides a principled connection between human interpretations of images and lightweight, fast image processing methods. Moving forward, it is critical that the community explicitly exploits such approaches, as the social image collections that we tackle, continue to grow larger.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Intentional framing; human interpretation of images; user intention; image classification

1. INTRODUCTION

Conventionally, multimedia researchers assume that what an image is about is primarily related to its literal subject matter, i.e., the visually depicted entities, events or scenes.

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2654894.

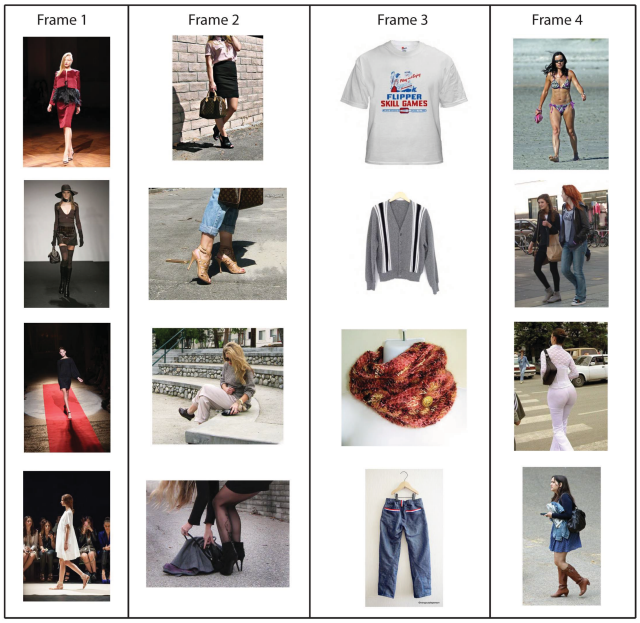

Figure 1: Four intentional frames reflect four different photographer intents (i.e., *how* users took the images). They are all indicative of the overall topic 'fashion' (i.e., *what* the images depict). Global-feature similarity suffices to capture intentional frames. Computationally intensive approaches are not necessarily required.

In this paper, we go beyond this conventional viewpoint and demonstrate that *what* an image is about is also reflected in *how* that image was taken. This new perspective benefits content-based approaches to large-scale social image collections, since it can be exploited in a simple, computationally lightweight fashion.

The core of the new perspective is the principle that we refer to as the *intentional framing*. We define intentional framing as, 'the sum of the choices made by photographers on exactly *how* to portray the subject matter that they have decided to photograph.' Note that intentional framing is a photographic act carried out by a photographer. Automatically captured images, such as security camera images, are not expected to exhibit framing effects. Fig. 1 provides an illustration of four ways in which 'fashion' is depicted in user photos on Flickr, a large online social photo-sharing community. These four different cases of *how* the subject matter of a photo is portrayed, correspond to four visually distinctive

intentional frames. In Sections 2 and 4, we will return to discuss this figure in more detail, including the relationship between these intentional frames and image similarly calculated automatically on the basis of global image features. Here, we first focus on introducing intentional framing, and describing its importance for multimedia research.

Our definition of intentional framing arises from the following considerations. When taking a photograph, the photographer does not click the shutter randomly, but first decides on a message and a subject. The decision process involves applying, either consciously or unconsciously, a set of conventions. These conventions can be thought as a recipe for a certain kind of image. This recipe is the intentional frame. The fact that the photographer applies a specific intentional frame leads to the generation of an image with distinguishable characteristics. These characteristics are visible in the image, and are used, again, either consciously or unconsciously, by humans in order to interpret the image.

When human viewers interpret an image at the level of its intentional frame, they are making a very high-level semantic judgement. The important role that intentional frame judgements play in human interpretations of images can be illustrated using a short thought experiment. Imagine a home with portraits of family members hanging on the wall. The subject matter (i.e., the *what*) of a portrait image is a person. In taking the image (i.e., the *how*), the photographer had the intent of creating a portrait. What would happen if the portraits of the family members were replaced with mugshots of the family members? This would be a strange situation. A visitor to this home would not easily be able to interpret the wall. A mugshot, like a portrait, portrays a person (i.e., the subject matter has remained the same). However, it is a very different image. The photographer who captures a mugshot has the intention of taking a picture that will be used by the police for identification purposes. This thought experiment demonstrates that two photos with the same literal subject matter (n.B. both a portrait and a mugshot are a photo of a person) are interpreted by the human mind in radically different ways.

The distinction between 'portrait' and 'mugshot' is a simple example used for the purposes of illustration. In this paper, we will investigate neither portraits nor mugshots specifically, but rather use a data-driven approach to explore intentional framing effects in large social image collections. However, by conducting this thought experiment, it is already possible to appreciate the profound implications of the concept of intentional framing for the multimedia field.

First, in order to arrive at image analysis algorithms that are truly capable of mimicking human image interpretation, image analysis algorithms should be 'aware' of the photographer's intention. In other words, they should be able to capture the visual differences that characterize images that were taken with different intents. For example, if humans find the difference between the intentional frames 'mugshot' and 'portrait' to be important, multimedia analysis algorithms need to make this distinction, too.

Second, sensitivity to very high-level semantic judgements, such as those related to intentional frames, will become critical as social image collections continue to grow larger. As pointed out by [13], an image retrieval system that indexes images by detecting basic concepts such as 'dog' cannot effectively support users to search huge social image collections. Even if the relative number of images depicting a 'dog' in the collection is small, if the collection is large enough, a 'dog' detector will detect thousands of dog images, i.e., many more than a user can use in a results' list. Instead, image analysis algorithms are needed which focus on specific aspects of images that are important to users and go beyond the basic concepts they depict. We do not claim that intentional framing is the only way in which human interpretations transcend the literal subject matter of an image. However, it is clear that it is an important contributing factor, and should for this reason be taken into account.

Finally, because of the fact that intentional framing impacts the overall 'look and feel' of images, differences in intentional framing can be captured by simple, lightweight approaches that exploit global representations. Such approaches are critical for allowing image analysis algorithms to scale and handle more and more images, as techniques for image indexing and retrieval are needed for larger and larger collections.

In short, intentional framing is important to the multimedia research community because it provides a principled motivation for applying lightweight approaches, exploiting global-feature representations to large-scale social image collections. The purpose of this paper is to establish the existence of intentional framing as a fundamental principle of human image interpretation, and to demonstrate its importance for content-based approaches to large-scale collections of social images. This paper makes three major contributions: (i) introduce intentional framing as a fundamental principle important for human interpretation of images at a high level of semantic abstraction, (ii) demonstrate that human-perceived similarity with respect to intentional frame corresponds to automatic similarity computed using global-feature representations of images and (iii) show that adopting the intentional framing perspective leads to a back-to-the-basics approach that relies exclusively on global-features to capture image semantics. Our approach delivers image classification rates that compete with the state of the art, while saving significantly in computational complexity.

We finish this section with an overview of the line of argumentation followed in the remainder of the paper. In Section 2, we discuss the related work, and demonstrate that although intentional framing is related to other phenomena studied in the literature, it cannot be reduced to any of them. In Section 3, we explain the concept of intentional framing in greater detail and provide illustrative examples.

Next, Section 4 presents two analysis experiments on human interpretations of images with respect to intentional framing. The experiments involve a user study and explore the judgments that humans make about images at the level of intentional frames. They establish the existence of a correspondence between automatic image similarity calculated on the basis of global features and human perceptions of images with respect to intentional frame.

This correspondence motivates us, in Section 5, to propose a back-to-the-basics simple search approach (SimSea) that leverages global feature representations to classify social images. We report results on standard image-classification task, 2013 Yahoo! Large-scale Flickr-tag Image Classification ACM Multimedia Grand Challenge. The results are surprising: a simple global-feature approach such as SimSea is able to compete with more 'sophisticated' content-based algorithms. Intentional framing, however, constitutes a principled reason why we should actually expect such re-

sults. Our conclusion, presented Section 6, opens up a future perspective.

2. RELATED WORK

Our coverage of related work first positions intentional framing with respect to other phenomena related to image semantics. We point to previous work that, plausibly, has taken advantage of the principle of intentional framing, without being aware of its existence. Finally, we cover the lightweight, search-based image classification approaches.

2.1 Intentional Frames and Image Semantics

Intentional framing is distinct from other aspects of image semantics because it focuses on 'how' the photographer has realized an image rather than 'what' is depicted in the image. There are three major research areas that seek to analyze images in terms of 'what' they portray: concepts, scenes, and events, which all focus on the literally depicted subject material of an image. Here, we cover each in turn.

Concepts: In context of image retrieval and analysis, concepts are objects and other entities that are literally depicted in images. The larger notion of a 'concept' derives from psychology and cognitive science, which has put forth various theories on human concept representations, e.g., concepts as definitions vs. concepts as mental images [33]. Independently of the exact mechanism involved, it is clear that concepts play a role in how humans store, organize and manipulate information about the world around them. For this reason, image analysis research has invested a great deal of effort into developing algorithms capable of detecting visual concepts [10].

An example of the variety covered by concepts is provided by the *ImageCLEF* concept detection task. Here, both categories of high-level semantic abstraction, such as 'fauna', 'age' or 'weather', are used alongside categories of lower abstraction levels, such as 'cat' and 'plant' [1]. Essentially, anything that is nameable by human observer can be considered a concept. Under this perspective, a scene or an event is considered a concept—scenes emphasize the positioning of elements and events include temporal sequence [23]. We turn to discuss both scenes and events in more detail.

Scenes: Scene interpretation has its roots in perception psychology. Scene perception describes the visual perception of an environment as seen by an observer at a given time. Rensink [27] describes perception of a scene as high, mid and low level processing steps. Long-term human learning results in a *scene schema* that interlinks the types of objects that occur together.

In the area of machine learning and content-based image retrieval, the notion of *gist* has been used to address the analysis of scenes. Gist originated in language analysis, and was introduced into image analysis by Friedman [7]. In the context of images, the gist is a description of a scene's overall meaning, such as 'farmyard', 'shopping center', or 'city' [27]. Olivia et al. [24] used global features to detect the gist of a scene. Global features capture global attributes of an image related to edges, colors or texture. Hays et al. [8, 9] used the idea of gist to address tasks related to geo-coordinates, such as geo-location detection and geo-scene completion.

The gist of a scene and the intentional framing of an image are related in the way that they both aim to capture global image characteristics. For this reason, global feature representations are suitable for both. However, intentional framing is a much broader notion that gist, since gist is restricted to 'what' is depicted in *scenes*, and intentional framing encompasses 'how' the subject material is presented in a *general social image*. The difference between scenes and intentional frames can be appreciated by considering Fig. 1. The notion of scenes is not adequate to account for the difference between the four frames. Instead we introduce intentional framing to go beyond the gist of scenes and to capture these differences.

Events: An event is a specific incident taking place at or over a given time span, involving one or more actors or objects and a specific place. Events often provide subject material for social images: weddings, parties, concerts, and sports events are favorite subjects of photos that users take and share online. Specific to the area of image analysis, an image may depict an event, but it is usually just a snapshot of the event and cannot cover every single aspect of it [12].

Work that has been done on Multimedia Event Detection, exemplified by the work in [22], is devoted to the detection of specific types of human activities, corresponding to types of events. This work focuses on detecting instances of particular event types, e.g., identify multimedia content that depicts a 'kiss' as a human activity. In contrast, a newer breed of work done on Social Event Detection is devoted to detect multimedia content that depicts a specific social event. This work focuses on identifying, for example, whether a photo was captured at a particular wedding. Examples of work on social events include [28], which uses candidate-retrieval methods and machine learning functions to automatically detect events in a stream and [26], which tackles social event detection by using multi-modal clustering and the integration of supervisory signals. Image analysis aiming to identify events, does not cover the same range of phenomena addressed by intentional framing. Note that although human activities and events are depicted in the images in Fig. 1, they do not provide a complete characterization of the differences between the four intentional frames.

To sum up, intentional framing, which focuses on 'how' images are taken, plays a significant role in human image interpretation. This role goes beyond aspects of image semantics that focus on 'what' is depicted in images, including concepts, scenes, and events. We close by mentioning a additional difference between 'the how' and 'the what' of images. Users/viewers recognize that two pictures are similar with respect to an intentional frame—referring again to Fig. 1, note the similarities among the images in each column. This recognition does not imply that it is easy, or even possible, to give an intentional frame a specific name. In contrast, concepts, scenes and events are often readily nameable (e.g., 'cat', 'farmyard' and 'kiss' above). We find that the fact that intentional frames are so difficult to be named, helps to explain why this important principle has been overlooked by the multimedia community thus far. This paper aims to compensate for past inattention.

2.2 Covert Exploitation of Framing

Although our basic position is that intentional framing has been overlooked by the multimedia community, we do *not* claim that it has never before been exploited. In this paper, we make the case that intentional framing is an integral part of the act of creating a photo and that, for this reason, we should expect the visual reflexes of intentional framing to act as a social signal that gives rise to exploitable pat-

terns in large collections of social images. If we consider intentional framing to be a fundamental principle underlying human image interpretation, then it is odd to assume that the multimedia community has entirely missed its existence up until this moment. Instead, we consider it to be highly likely that past work in the area of image analysis and retrieval has made use intentional framing effects without being aware of it. Specifically, we make the point that any approach that exploits content based comparison, e.g., pairwise similarity, of images may also be capturing regularities in 'how' the images were photographed alongside of regularities in 'what' the images depict.

Here we mention a two specific social image analysis approaches that we suspect might already exploiting 'how' alongside of 'what'. In Li et al. [15], social tag relevance is learned with a visual neighbor voting algorithm. The approach searches for similar images based on a query image. It cannot be excluded that such similarity is indirectly picking up on 'how' images are taken, in addition to 'what' they literally depict. Another example is Liu et al.'s [16] work on tag propagation. This work makes use of a tag-specific visual sub-vocabulary. Such a sub-vocabulary could easily be exploiting 'how' images are photographed alongside of 'what' they depict. We believe that there are a large number of examples of research that may be unwittingly exploiting intentional framing. An key contribution of this paper is to point out the existence of intentional framing, with the goal of stimulating research on its explicit exploitation. If an algorithm already benefits implicitly from intentional framing, we believe it can only be improved by understanding the extent of this benefit, and by actively seeking to enhance it. In this paper, we do not directly quantify the benefits of intentional framing, but rather focus on laying a solid groundwork for future work in that direction.

2.3 Search-based Image Classification

Finally, we turn to discussing work related to our search-based image classification approach, SimSea. We would like to explicitly point out that SimSea itself does not constitute a major contribution of this paper. Rather, we introduce SimSea as a back-to-the-basics algorithm that exploits global-feature representations. Its effectiveness is rather mysterious, until we take the perspective that global features are capable of capturing the semantics of large-scale social image collections because they are sensitive to the semantics associated with intentional frames.

SimSea is a search-based approach representing a variant of the well-known k-NN algorithm. The multimedia item to be classified is used as a query, and a similarity metric is applied to retrieve a ranked list of the most similar items in a collection of multimedia items that has been labeled with category labels. The category labels of the top-ranked items are then propagated to the query image. The work most closely related to ours is the geo-visual ranking approach to content-based prediction of image location [14]. Here, the location of a photo is predicted by using the photo as a query to retrieve a list of geo-visual neighbors from a social image collection, and propagating the most visual likely location to the photo.

Additionally, Wang et al. [34] and Yang and Hanjalic [35] use similar approaches in their work. However, these use both image features and text features, and focus on re-ranking search results. In contrast, our approach relies exclusively on the visual channel, is not deployed for concept detection, and is tested at a much larger scale.

3. INTENTIONAL FRAMING

In this section, we present the concept of framing in more detail. Specifically, we discuss photographers' choices that lead to intentional framing, and we provide examples of intentional framing in social image collections, illustrating its link to human interpretations and its generality.

In the most general sense, frames are organizational structures in which information is communicated or understood. They have been extensively studied in the field of communication, which investigates a wide and disparate range of framing phenomena [6, 29]. Across phenomena, however, it is agreed that frames regulate *how* information is communicated, rather than directly determining *what* is communicated. In describing how frames work, Entman [6] states, 'Frames highlight some bits of information about an item that is the subject of a communication, thereby elevating them in salience' (p. 53). Similarly, the decisions that a photographer makes when taking a photograph determine which information in the photographs gets noticed or interpreted as important by the viewer.

3.1 Photographers' Choices

Recall that we have defined intentional framing as, 'the sum of the choices made by photographers on exactly *how* to portray the subject matter that they have decided to photograph.' We use the term 'intentional framing', rather than simply 'framing' to emphasize that the 'frame' is the visible reflexes of the *intent* of the photographer to create a certain type of image. The term 'intentional framing' also disambiguates our use of the word 'framing' from another use common in photography. Specifically, photographers use 'framing' to refer to positioning the subject of a photo within a door or other opening that acts like a window frame in a photograph. This sense of 'framing' is not the one that we are addressing here.

Choices a photographer makes to achieve certain types of framing include color distribution, lighting, positions of objects and people, camera angle, depth of field, and focus. They also include the choice of the precise moment during ongoing action at which the image is shot. In this way, the photographer also influences exactly what is depicted in the image, for example, facial expressions of the people appearing in the image. In general, the influence of the photographer reflects not so much personal choices, but rather shared expectations between photographers and viewers about how photos portray the world. These expectations constitute a set of conventions that allow viewers to interpret photos. Radically creative photography may make breaks with conventions, but photos that stray too far from familiar recipes are difficult to interpret.

The importance of intentional framing for photographers is witnessed by the way how it is described on websites that teach photography. For example, Fodors provides a web tutorial for travel photography [31]. Several different methods for framing photos are described, each related to different subject matter: 'classic vacation shots', 'the man-made world', 'the natural world', 'the elements' and 'people'. Each is broken down into finer-grained topic related categories.

Clearly, the decisions that photographers make that determine intentional framing are closely related to composi-

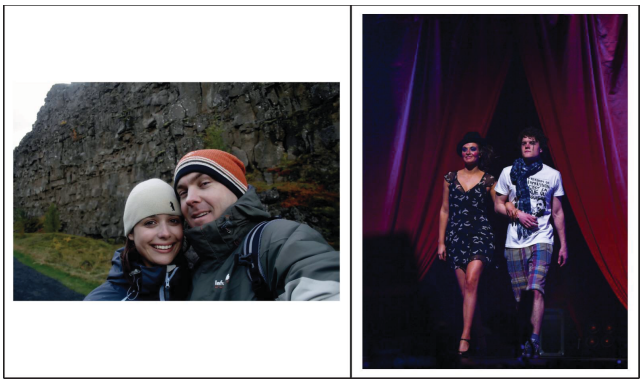

Figure 2: Example Flickr images that depict the same visual contents (a woman and a man), but correspond to two different intentional frames (left: holiday memories and right: fashion)

Table 1: The metadata for the images in Fig. 2 reflecting the different underlying intents of the user/photographer

	Left image	Right image
Title	Thingvellir	DCU Fashion Show 2009
Tags	trip, iceland, reykjavik, ...	fashion, Cirque Du Couture dcu, Couture, ...
Description	Iceland 2009.	DCU Style Society presents DCU FASHION SHOW 2009 ...

tion. The composition of a photograph includes the arrangement of objects, the angle, the focus or the distribution of colors in a photograph. Although composition choices contribute to intentional framing, intentional framing cannot be reduced to composition. We explicitly point out that intentional framing also includes choices beyond composition, such as whether human subjects are visibly expressing emotion, and the choice of the exact setting. The fact that photographers consider intentional framing as a way to extend beyond composition is illustrated by the organization of the tutorial [31]. Here, methods for framing photos are *not* treated under the heading 'Photography composition rules'. There are rather separate sections dedicated to composition and to framing. We are interested in the broader concept of framing rather than the narrower concept of composition as it is more tightly related to the topic of the image, which makes it a better indicator of image semantics.

3.2 Intentional Frames in Practice

In Fig. 2, we present two social images from Flickr that both depict the same basic content, a woman and a man, but differ in respect to their intentional framing. We can gain insight into the intent of the users that took these images by inspecting their titles, tags and descriptions, shown in Tab. 1. This metadata leads to the conclusion that the intent behind the image on the left is to capture a memory of a trip and the intent behind the image on the right is to depict fashion.

This difference in intent can also be seen in how the users who took the photos have chose to frame them. Although both images show a man and a woman, in the image on the left, the user has chose to make a 'selfie' in an outdoor setting that focuses on faces and smiles, and in the image on the right, the setting is an illuminated stage and the focus is on the clothing. An inspection of Fig. 2 reveals that these choices have visual reflexes in the photos. The visual manifestations of intentional framing signal to the viewer that one photo should be interpreted as representing holiday memories and the other as depicting fashion.

With this example we would also like to stress the point that textual metadata could possibly help in the differentiation of photos on the basis of their intentional frame. Our focus here, however, is on visually observable intentional framing effects and on content-based approaches. For this reason, we do not consider textual features any further.

3.3 Viewers' Interpretations

The study of framing has its roots in the field of social psychology, where a frame describes a general, mainly subliminal, basic idea at play during perception or interpretation. The notion of 'frame' is thus tightly related to *Gestalt*, the perception of the essence or shape of an entity's complete form [11]. Specific to image perception is the notion of 'gist' [7], i.e., what is perceived from an image at a glance. We have already noted that gist-based methods have been applied by multimedia researchers to the problem of analyzing images that depict scenes.

Viewer interpretations of images are tightly synchronized with the intentional framing that is chosen by a photographer. In fact, the intentional framing of the image constitutes a signal from the photographer to the viewer about how the image should be interpreted. For some subject material, photographers often use highly conventionalized intentional frames. For example, nearly everyone can bring a standard picture in mind of how a traditional bridal pair appears in a wedding photo, or a how a public figure is represented in a certain role, e.g., a politician delivering an inspiring speech.

For other subject material, the intentional framing is less tightly linked to the subject matter, but rather more closely related to the underlying goal or purpose. For example, [18] establishes a typology of photographer intentions. This work demonstrates the reasons for which people take social images range from sharing emotions to recalling a feeling or collecting and storing information.

Our work does not depend on explicitly identifying or cataloguing intentional frames corresponding to all possible image topics, or photographer goals and purposes. We are rather interested in the fact that photographers use intentional frames to create photos, and that users/viewers differentiate photos on the basis of intentional frames. In other words, our work is focused on establishing that, alongside of *what* photos depict, *how* photos are taken is important for human interpretation of image semantics.

We point out that intentional framing is closely linked to the notion of *connotation*. In the area of images, connotation refers to those aspects of image interpretation that go beyond the literally depicted subject material of the image. In his seminal essay in [2], *The Photographic Message*, Roland Barthes characterizes connotation as 'the imposition of second meaning on the photographic proper' (p. 20). Intentional framing can also be considered a 'second meaning'. However, understanding connotation involves interpreting 'what' is depicted in an image. For example, red roses are commonly considered to have connotations of love. In contrast, intentional framing keeps the focus specifically on 'how' image content is depicted, and the way that photog-

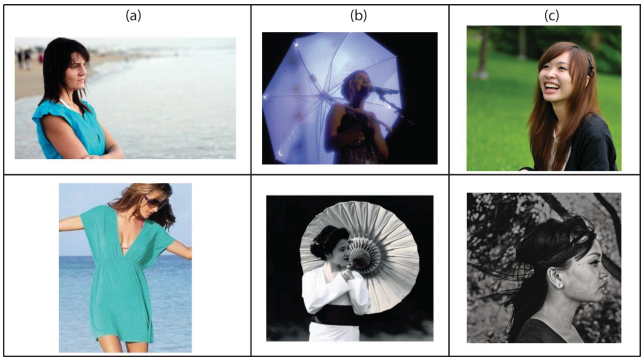

Figure 3: Pairs of photos that contrast with respect to intentional frame. The pairs differ with respect to the interpretations of human viewers, impressionistically described as: (a) feeling vs. fashion, (b) performance vs. history (c) personality vs. art

rapher choices related to 'how' are reflected in visual characteristics of an image. We remark that ultimately intentional framing may lead the multimedia research community to more effective exploitation of overall image connotations.

3.4 Generality of Intentional Framing

Intentional framing is a general phenomenon underlying social images. The visual reflexes of photographers' intent constitute a social signal that influences the global patterns that exist in a large collection of social images. Some photographers might take images unthinkingly, but most conceptualize their images to at least a minimum extent. Photographer choices, in turn, impact exactly *how* the subject material is depicted in the image. We do not claim that intentional framing constitutes a strong signal within a social image collection. Rather, our position is that this signal exists, and that it is strong enough to be effectively exploited. Here, we present additional examples to demonstrate that intentional framing takes different forms, and that a large range of images can be differentiated on the basis of intentional framing.

The Flickr images in Fig. 3 are arranged in pairs that differ with respect to intentional framing. The contrast between the two photos in each pair demonstrates that if two photos depicting the same concept or entity use different intentional frames, the result is two images with different interpretations. Consider the two photos in column (a). The top photo is about what the woman in the blue dress is feeling, and the bottom photo is about the blue dress. The contrast is due to the framing choices made by the photographers who took these photos. These choices include not only the ratio of the frame filled by the dress vs. the ratio filled by the background water, but also with the depth of field, the overall color palette, and the emotion projected by the human subject, and the subject's posture. In other words, it is intentional framing and not the depicted visual concepts that serve to distinguish these images from the point of view of a human interpreter.

Similar observations can be made about the photo pairs in (b) and (c). In (b), one set of photographer choices leads to an image depicting an ongoing performance (top), and the other to an image that documents history (bottom). Note that these two photos are very similar with respect to their basic composition, but different in their interpretation. This pair illustrates how intentional framing includes, but goes beyond, photographers' composition choices. In (c), one set of photographer choices leads to an image that conveys the happy personality of the subject (top) and another set of choices lead to a photo with a somber mood (bottom) that could be considered a work of art, more than a testimony to the personality of the person displayed.

It is important to note that the descriptions we use to refer to viewer interpretations of framing are impressionistic. We do *not* claim that these are the only possible descriptions, or that algorithms should predict these interpretations directly. Our point is that intentional framing is important for human image interpretations, and content-based algorithms should not be 'blind' to its existence. The larger message is that multimedia researchers should not indiscriminately assume that content-based image methods must 'compensate' for the visual variability of depicted objects, scenes, and images. Such approaches will lead to image analysis and retrieval systems that cannot possibly be sensitive to the difference in human interpretation between the pair of images in (a). Instead we advocate systems that admit the possibility that differences important for human semantic interpretation of images are related to intentional frames.

4. HUMAN VIEWS ON FRAMING

In this section, we empirically investigate the phenomenon of intentional framing. On the basis of the discussion in Section 3, we expect intentional framing to manifest itself in a collection of social images in the form of clusters of images that are homogenous in terms of their overall 'look and feel'. For this reason, we study clusters of images that are created automatically using global feature representations. We are interested in two aspects of these clusters, which we investigate in two analysis experiments involving human judgments collected via user studies.

The first experiment explores the correspondence of global-feature clusters with human judgements of photographer intent. The second experiment explores the correspondence of global-feature clusters with image semantics in the form of a higher level topic, in this case, 'fashion'. Each experiment consists of two steps, first, the *clustering step*, in which we create clusters in a social image collection, and, second, the *correlation step* in which we analyze the relationship between the clusters and human judgements related to intentional framing.

4.1 Global Features and Photographer Intent

According to the principle of intentional framing, the intent of the photographer guides the decisions made by the photographer while conceptualizing an image, resulting in an image with a particular intentional frame. However, since intentional framing results from a general recipe for a photograph, rather than specific rules, and, since photographers apply this recipe only to varying degrees, we, yet, know nothing about the visual variability that characterizes intentional frames. For this reason, the goal of our first experiment, is to demonstrate that it is indeed conceivable that global features can capture the regularities of frames.

For this experiment, we use the *Photo Intentions* data set that has been created by Lux et al. [20], and consists of 1,310 Flickr photos annotated with *photographer intent categories*. The categories correspond to general photographer goals in taking a picture: (i) *preserve a good feeling*, (ii) *preserve a*

bad feeling, (iii) *show it to family and friends*, (iv) *publish it on-line*, (v) *support a task of mine* and (vi) *recall a specific situation*, and were chosen on the basis of a previous user study carried out by [18]. The images in the data set were annotated by the users who took them, who were contacted by Lux et al. [20] via Flickr. The category labels provided by the photographers were verified using a crowdsourcing experiment carried out on Mechanical Turk. As explained in detail in [20], five crowdworkers judged each image, and rated it with respect to each of the six intent categories. These ratings, used in our experiment, reflected the association of the image with each of the six categories using a 5-point Likert scale.

It is important to note that in this experiment, we do not assume that the photographer's intent category corresponds just to one single intentional frame. Instead, we take these categories to involve multiple closely related intentional frames that photographers use to accomplish a particular goal or purpose. We assume that if visual clusters correspond to intentional frames, then they will also be correlated with intent categories that encompass multiple frames. To our knowledge, our data set is the largest publicly available collection of social images that includes information about the intent of the photographer.

The *clustering step* in our experiment was carried out as follows. A selection of common global features was made, and the features were extracted from the images using *LIRE* (latest version[1]) [19]. For each type of global feature, clustering is performed using *Weka* (version 3[2]). We chose X-means clustering [25], since it determines the number of the clusters automatically, which is important for the experiment.

The *correlation step* was carried out for each different global-featuring clustering of the images. The purpose of the correlation step was to compare the visual closeness of the images in a cluster, with the human perception of whether the images were 'close' with respect to the intent of the photographer. We analyze each global feature clustering with respect to each intent category separately. Specifically, we calculate the Pearson correlation between the mean square distance of the images in a cluster from their cluster centroid and the mean of the Likert-scale ratings reflecting the degree to which the images in the cluster are associated with the intent category.

Tab. 2 summarizes the results, and demonstrates that the experiment uncovered the existence of a number of cases in which the tightness of visual clusters correlates (> 0.3) with human agreement on the photographer's intent.

These cases (n.B. they are negative correlations) are indicated with black. It can be seen that certain features seem to be particularly well suited for certain categories. For example, FCTH is the best feature for detecting photos for which the photographer's intent was *publish on-line* (i.e., in a blog). It is important to note that the results of this experiment must be seen as a *suggestion* that global features can capture photographer intent. There are also cases of positive correlation, which are marked in white, where it is clear that other effects are at play. However, if there was no relationship between global features and photographer intent, we would have expected a table that was entirely grey,

[1]https://code.google.com/p/lire/source/checkout
[2]http://www.cs.waikato.ac.nz/ml/weka/downloading.html

Table 2: Correlation of global-feature-based clusters (MSE) and human agreement on photographer intent on the *Photo Intentions* data set.

Feature	recall situation	preserve good feeling	publish online	show to family and friends	support task	preserve bad feeling
CEDD						
FCTH						
Gabor						
Tamura						
Luminance Layout						
Scaleable Color						
Opponent Histogram						
AutoColor Correlogram						
JPEG Coefficent						
Edge Histogram						
PHOG						
JCD						
JointHistogram						

which is not the case. Encouraged by this initial experiment, we turned to a second, larger-scale experiment, that investigates the connection between image clusters and topical semantics.

4.2 Intentional Framing and Topic

Our second analysis experiment is closely related to the title of this paper. It investigates the connection between 'how' an image is taken and 'what' that image depicts. Recall that the intentional frame that a photographer chooses is related to the particular subject material that is portrayed in the image (i.e., the topic). The importance of the relationship between intentional framing and topic is the following: if visual patterns of intentional framing in social data sets are indicative of topic, then they can be exploited for content-based tasks such as analysis of image semantics, and image retrieval.

For this experiment we use the *Fashion 10000* data set for Flickr images, which was created by Loni et al. [17] for the purpose of developing classifiers to detect fashion images in social image collections. The data set consists of 30,000 images and was collected to contain a significant portion of images (>10,000 of them) that are related to fashion and clothing. Further details on how the +fashion/-fashion labels were generated can be found in [17].

On the basis of the results of the previous experiment, we expect that clustering using global feature representations are indeed capable of picking up visual regularities in the data related to intentional framing. This experiment had the goal of uncovering a relationship between the visual tightness of clusters and human judgements that these clusters were related to the overall topic of fashion.

Because the *Fashion 10000* data set is an order of magnitude larger than the intention data set, we first carried out clustering, and then submitted the clusters to a group of human judges. The *clustering step* in this experiment was carried out by first determining an optimal global feature representation for the data using average information gain. Under this assumption, the following features were identified as useful for the data set: CEDD, FCTH, JCD, PHOG, ColorLayout, JPEG coefficient histogram and ScalableColor [19]. As before, X-means clustering was applied, resulting in 62 clusters.

In the *correlation step*, a set of human subjects were presented with 62 screens of images, each screen containing im-

ages sampled from one of the 62 clusters (n.B. the clusters were too large to judge in their entirety). In total, there were 10 participants who judged the images. The participants were selected by convenience sampling from the immediate environment of the authors, and briefed to ensure that they had an adequate understanding of social fashion. The participants then judged the consistency of the clusters with respect to fashion.

As in the previous experiment, we calculate the Pearson correlation between the mean square distance of the images in a cluster from the cluster's centroid (reflecting visual tightness of the cluster) and the average human agreement about the fact that the cluster reflects fashion. The result was a negative correlation, -0.56 (i.e., small, tight clusters are associated with clear human judgements of topic focus). This result provides evidence that global features are indeed able to build clusters that partition fashion from non-fashion images in the data set. This experiment supports the conclusion that 'how' images are taken, as reflected in global-feature-based clustering, is indicative of the topical subject matter that they contain.

We conclude this section by mentioning that qualitatively the outcomes of the second experiment were striking. Specifically, the four groups of intentional frames in Fig. 1 in Section 1 were *not* hand selected from the data. Rather, these four intentional frames represent clusters that were formed using global-feature-based clustering in the second analysis experiment. These clusters serve as a compelling illustration of the link between global features and semantic image content. Our position is that this link exists, because of the photographer's tendency to take pictures of specific content which follows a set of intentional frames. This effect is stable enough to be a useful visual signal within large-scale social image collections.

5. CONTENT-BASED CLASSIFICATION

The evidence in Section 4 suggest that there is a link between global feature representations and image topic that is mediated by intentional framing. Motivated by this evidence, we carry out an experiment designed to exploit that link. The experiment involves large-scale classification of social images into tag-classes. Intentional framing gives us reason to believe that images belonging to a certain tag-class, and therefore containing certain topical subject material, will be characterized by patterns of intentional framing. These patterns reflect typical sets of choices made by users on 'how' to make a photograph that are related to the subject material that they are photographing. We do not expect the effects to be strong. Instead, our goal is to present plausible proof that the effects of intentional framing are exploitable for a task related to image semantics.

5.1 Data Set and Experimental Setup

We carry out our content-based classification experiment on the *Yahoo! Flickr Creative Common Images tagged with ten concepts, version 1.0 data set*[3] that was used for the 2013 Yahoo! Large-scale Flickr-tag Image Classification ACM Multimedia Grand Challenge. The data set consists of 1.5 million training images associated with ten equally-sized tag-classes and 500,000 test images. The tag-classes are: *2012*, *beach*, *food*, *london*, *music*, *nature*, *people*, *sky*, *travel*, and *wedding*. This data set is considered challenging not only due to its scale, but also because each topical tag-class is characterized by a very high degree of visual variability. Our choice of a standard data set allows us to compare our approach to the performance achieved by current state-of-the-art methods.

In order to make clear how the theory of intentional framing is expected to contribute to the performance of a classifier on such a classification task, we consider the class 'London' in more detail. Images taken all over London will be tagged 'London', giving rise to a high level of visual diversity. However, because we are looking specifically at social images, we expect that people are taking pictures of London mainly with the intention of documenting the city, for example, as tourists, as residents or as journalists. For this reason, we expect photos to be generally associated with key intentional frames, examples might be, cityscape photos, photos that emphasize a sense of place, and photos taken to preserve memories. These intentional frames are indicative of the topic 'London' the way that the four frames in Fig. 1 are indicative of the topic 'fashion'.

The intentional frames are not expected to be mutually exclusive among tag-classes. However, they are expected to support discrimination well enough to act as indicators of tag-classes. For example, it is reasonable to expect that more cityscape photos would be anticipated in the tag-class 'London' than in the tag-class 'Food'. In this way, intentional framing can be anticipated to deliver performance on this task—sensitivity to the specific literally depicted content of the images (i.e., detecting specific food items or specific city landmarks) is not necessary.

Our simple search classification approach, SimSea, is a variant of the k-NN algorithm. We choose a back-to-the-basics approach because of its computational simplicity and the speed that it delivers on large scale image classification problems. As previously mentioned, SimSea is not itself novel. Our novel contribution is that intentional framing provides a principled explanation as to why an algorithm such as SimSea should work well on a large collection of social images.

SimSea is implemented by extending the LIRE framework [19] with an implementation of a search based classifier. The following global features were used: JCD [5], CL [4], OH [32] and PHOG [3]. These features were selected using information gain calculated on the development set (a subset of the training set).

For retrieval we employ the inverted index strategy to index the hash values, like terms describing the actual image. To query the system, we create a term-based query from the hashes of the query image.

Classification proceeds as follows. Each search result of a given tag-class is counted as one vote in favor of assigning the query image to that tag-class. The class with the most votes wins, and is returned as the classified class for the query image. In case of a draw, the occurrences of the class is weighted by the rank of each image with the same class. The weight function is defined as:

$$c = \arg\max_{c \in C} \{ClassScore(c)\}$$

$$ClassScore(c) = |c|. \sum_{I_i \in \{I_i | Class(I_i) = c\}} rank(I_i)^{-1}$$

The class with the highest $ClassScore$ of all classes $c \in C$ is chosen as class c of the image. The $ClassScore$ is calculated

[3]see http://webscope.sandbox.yahoo.com

Table 3: iAP per class on the development set.

	JCD	CL	OH	PHOG
2012	0.198	0.128	0.13	0.104
beach	0.448	0.487	0.342	0.534
food	0.531	0.492	0.389	0.352
london	0.244	0.201	0.146	0.347
music	0.526	0.457	0.495	0.164
nature	0.502	0.41	0.435	0.503
people	0.264	0.227	0.244	0.105
sky	0.628	0.601	0.544	0.473
travel	0.139	0.101	0.128	0.112
wedding	0.463	0.272	0.262	0.235

by counting the occurrences of each class c and multiply it with the summed $WeightedRankScore$. $rank(I_i) : \{I_i\} \rightarrow \mathbb{N}$ gives the rank of an image. The $WeightedRankScore$ is the sum of $rank(I_i)^{-1}$ scores for each class. The search time of this approach is well below 300 ms for the 1,500,000 indexed images. Due to the nature of the global feature search, the search time will scale sub-linearly with the number of images in the index.

5.2 Experimental Results

For evaluation metric, we adopt mean interpolated averaged precision ($MiAP$), which was also used to evaluate the results of the Grand Challenge.

Our first step is to use the development set to determine optimal settings. Our development set is based on a sub-set of the training data that includes 1,000 randomly selected images per class. A setting of $k = 50$, was determined to be optimal. Tab. 3 reports results for $k = 50$ using a variety of global features. It can be seen that average precision is very similar for all choices of features. However, JCD achieves the highest overall mean interpolated average precision (0.417), and is therefore chosen for the experiment on the test set.

Our second step is to carry out classification on whole test data set (500k images; 50k images per tag-class). Applying the optimal settings determined on the development set, i.e., $k = 50$ and JCD features, we performed our second experiment with the test data set featuring 50k images per class which leads to a $MiAP$ over all classes of 0.391.

Tab. 4 shows the $MiAP$ over all classes for the 500k test set and all 1.5 million training images to train the model, compared with the best results of the ACMMM Grand Challenge 2013 from Mantziou et al. [21]. The best performing reported result from the second approach from Su et al. [30] is not compared because the reported $MiAP$ excludes the tag-class *2012*. However, they also use a concept detection approach, Hessian affine (Concept 1 (HA)), which is more comparable. The comparison shows that our *SimSea* approach which uses *only one* global feature, nearly can reach the performance of the other approaches and in one case, Concept 1 (HA) [30], can reach better performance.

As baseline we use the results of SMaL [21] (Local 1 (SMaL)) and SVM [21] (Local 2 (SVM)), which both rely on local features and complex learning algorithms, because they report the $MiAP$ for each class which makes it better comparable with our results. Based on the $MiAP$ We calculated the statistical significance (Wilcoxon Signed-Rank Test) with a significance level of 0.01. This leads to p-value of 0.5754 for SimSea versus Local 1 (SMaL) and a p-value of 0.3320 for SimSea versus Local 2 (SVM). This test shows that the difference is not statistically significant and therefore our methods performance cannot be interpreted as worse (or better) than the Local 1 and Local 2 approach. For the sake of completeness we mention that all approaches outperform the dominant class baseline, which is 0.1.

Table 4: SimSea vs. best results from the ACMMM Grand Challenge 2013. The difference with Local 1 and Local 2 is not statistically significant

	SimSea	Local 1 (SMaL) [21]	Local 2 (SVM) [21]	Concept 1 (HA) [30]
$MiAP$	0.391	0.422	0.413	0.37

It is important to point out that the run-time performance of our approach is very good. Classification of a single image takes roughly 300 ms on a current Windows 7, Intel Core i7, 16GB PC. This is faster than 10 minutes for Local 1 (SMaL) and 2.5 seconds for Local 2 (SVM) on a 24-core Intel Xeon Q6600 2.0Ghz with 128GB RAM reported in [21].

6. CONCLUSION AND OUTLOOK

This paper has presented a proof for the importance of visual patterns in large collections of social images that exist due to underlying consistencies in how photographers choose to make images. Conceptually, we have turned from considering an image as something that is viewed by a user ('what' is shown in the image) to considering an image to be something that was created by a photographer ('how' the image was captured by the photographer). This shift of perspective allows us open up a new set of commonalities between images. These commonalities are useful for image analysis and retrieval because they both connect images at the level of pixels, and also correspond to connections perceived by humans interpreting images.

Specifically, this paper has introduced intentional framing, a principle that accounts for the connection between the decisions made by photographers that are related to the subject material they photograph, and visual characteristics of social images. We show that global feature representations of images are connected to human perceptions of photographer intent, and also that intentional frames chosen by photographers are connected to the semantics of images at the level of topic

We report the results of a large-scale image classification experiment that makes use of a back-to-the-basics simple search approach exploiting global feature representations of images. If we consider that the perspective image topic is exclusively related to 'what' images predict, the good performance of this simple approach is mysterious. Global features are known to be related to the overall 'look and feel' of images and not necessarily to specific semantic content. However, once we understand that intentional framing has the ability to act as a bridge between global characteristics of an image, and interpretations of image semantics, these results are expected.

The principled account that intentional framing provides for the results of our content-based classification experiment opens a new vista for multimedia research. We consider the results of this experiment to reflect the force of intentional framing at work in a large-scale social image collection. However, it measures that force, at best, only indirectly. Additional work is required to understand the exact nature of that force, the factors that influence it, and how it can best be harnessed in the service of image analysis and retrieval. For example, we would not expect an intentional-

framing-sensitive approach to work well on a collection of images not taken by human photographers, e.g., Google street view images. Such a collection would have only a single robotically created frame, and image semantics would not be differentiated at this level. Further experimentation is necessary to test this hypothesis and to discover how to exploit the intentional framing principle to its full potential.

Taken as a whole, the evidence presented in this paper serves to demonstrate the high promise of the 'how' perspective for social images. We conclude that intentional framing opens the possibility of the exploitation of lightweight approaches that contribute to the development of a new breed of fast image analysis and content-based retrieval algorithms for large-scale social image collections.

7. ACKNOWLEDGMENTS

We would like to thank Xinchao Li, Yue Shi and Oge Marques for fruitful discussion of framing. M. Riegler is a recipent of the Excellence Scholarship of the Industrialists' association Carinthia. which provided partial support for this research. Support was also received from Lakeside Labs GmbH, Klagenfurt, Austria, the European Regional Development Fund, the Carinthian Economic Promotion Fund (KWF-20214/25557/37319), EC FP7 project CUbRIK (287704), the iAD center for Research-based Innovation (174867) funded by the Norwegian Research Council, and the Dutch national program COMMIT. C. Kofler is a recipient of the Google Europe Doctoral Fellowship in Video Search, which also provided partial support.

8. REFERENCES

[1] ImageCLEF Task. http://imageclef.org/2012. [lv., 08, 13].

[2] R. Barthes. *Image Music Text*. Hill and Wang, 1977.

[3] A. Bosch, A. Zisserman, and X. Munoz. Representing Shape with a Spatial Pyramid Kernel. In *Proceedings of the CIVR '07*, pages 401–408, New York, NY, USA, 2007.

[4] S.-F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):688–695, June 2001.

[5] S. A. Chatzichristofis, Y. S. Boutalis, and M. Lux. Selection of the Proper Compact Composite Descriptor for Improving Content Based Image Retrieval. In *Proceedings of the SPPRA 09'*, 2009.

[6] R. M. Entman. Framing: Toward Clarification of a Fractured Paradigm. *Journal of communication*, 43(4):51–58, 1993.

[7] A. Friedman. Framing Pictures: The Role of Knowledge in Automatized Encoding and Memory for Gist. *Journal of experimental psychology: General*, 108(3):316, 1979.

[8] J. Hays and A. A. Efros. IM2GPS: Estimating Geographic Information from a Single Image. In *Proceedings of the CVPR 08'*, pages 1–8.

[9] J. Hays and A. A. Efros. Scene Completion Using Millions of Photographs. In *Proceedings of the ACM TOG 07'*, volume 26, page 4, 2007.

[10] M. J. Huiskes, B. Thomee, and M. S. Lew. New Trends and Ideas in Visual Concept Detection: The MIR Flickr Retrieval Evaluation Initiative. In *Proceedings of the ACM ICMR 10'*, pages 527–536, 2010.

[11] G. Humphrey. The Psychology of the Gestalt. *Journal of Educational Psychology*, 15(7):401, 1924.

[12] J. Kim. Events as Property Exemplifications. In *Action theory*, pages 159–177. Springer, 1976.

[13] M. Larson, M. Melenhorst, M. Menéndez, and P. Xu. Using Crowdsourcing to Capture Complexity in Human Interpretations of Multimedia Content. In *Fusion in Computer Vision*, pages 229–269. Springer, 2014.

[14] X. Li, M. Larson, and A. Hanjalic. Geo-visual Ranking for Location Prediction of Social Images. In *Proceedings of ACM ICMR 13'*, pages 81–88. ACM, 2013.

[15] X. Li, C. G. Snoek, and M. Worring. Learning Social Tag Relevance by Neighbor Voting. *Multimedia, IEEE Transactions on*, 11(7):1310–1322, 2009.

[16] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang. Image Retagging Using Collaborative Tag Propagation. *Multimedia, IEEE Transactions on*, 13(4):702–712, 2011.

[17] B. Loni, L. Y. Cheung, M. Riegler, A. Bozzon, L. Gottlieb, and M. Larson. Fashion 10000: An Enriched Social Image Dataset for Fashion and Clothing. In *Proceedings of ACM MMSys 14'*, pages 41–46, New York, NY, USA, 2014. ACM.

[18] M. Lux, M. Kogler, and M. del Fabro. Why Did You Take This Photo: A Study on User Intentions in Digital Photo Productions. In *Proceedings of SAPMIA 10'*, SAPMIA '10, pages 41–44, New York, NY, USA, 2010. ACM.

[19] M. Lux and O. Marques. Visual Information Retrieval Using Java and LIRE. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 5(1):1–112, 2013.

[20] M. Lux, M. Taschwer, and O. Marques. A Closer Look at Photographers' Intentions: A Test Dataset. In *Proceedings of the ACM CrowdMM 12'*, pages 17–18. ACM, 2012.

[21] E. Mantziou, S. Papadopoulos, and Y. Kompatsiaris. Scalable Training with Approximate Incremental Laplacian Eigenmaps and PCA. In *Proceedings of the ACM MM 13'*, pages 381–384, 2013.

[22] T. B. Moeslund, O. Javed, Y.-G. Jiang, and R. Manmatha. Special Issue on Multimedia Event Detection. *Mach. Vision Appl.*, 25(1):1–4, Jan. 2014.

[23] M. Naphade, J. R. Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale Concept Ontology for Multimedia. *MM IEEE*, pages 86–91, 2006.

[24] A. Oliva and A. Torralba. Building the Gist of a Scene: The Role of Global Image Features in Recognition. *Progress in brain research*, 155:23–36, 2006.

[25] D. Pelleg, A. W. Moore, et al. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of ICML 00'*, pages 727–734, 2000.

[26] G. Petkos, S. Papadopoulos, and Y. Kompatsiaris. Social Event Detection Using Multimodal Clustering and Integrating Supervisory Signals. In *Proceedings of ACM ICMR 12'*, page 23, 2012.

[27] R. A. Rensink. Scene Perception. 7:151–155, 2000.

[28] T. Reuter and P. Cimiano. Event-based Classification of Social Media Streams. In *Proceedings of ACM ICMR 12'*, page 22, 2012.

[29] D. A. Scheufele. Framing as a Theory of Media Effects. *Journal of communication*, 49(1):103–122, 1999.

[30] Y.-C. Su, T.-H. Chiu, G.-L. Wu, C.-Y. Yeh, F. Wu, and W. Hsu. Flickr-tag Prediction Using Multi-modal Fusion and Meta Information. In *Proceedings of ACM MM 13'*, pages 353–356, 2013.

[31] F. Travel. Framing of Images from Photographers View. http://www.fodors.com/travel-photography/. [lv., 11, 13].

[32] K. E. van de Sande, T. Gevers, and C. G. Snoek. Evaluating Color Descriptors for Object and Scene Recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596, 2010.

[33] S. Vinner. Concept Definition, Concept Image and the Notion of Function. *International Journal of Mathematical Education in Science and Technology*, 14(3):293–305, 1983.

[34] L. Wang, L. Yang, and X. Tian. Query Aware Visual Similarity Propagation for Image Search Reranking. In *Proceedings of ACM MM 09'*, pages 725–728. ACM, 2009.

[35] L. Yang and A. Hanjalic. Supervised Reranking for Web Image Search. In *Proceedings of ACM MM 10'*, pages 183–192. ACM, 2010.