A Reformulation of the Semantic Gap Problem in Content-Based Image Retrieval Scenarios

Tommaso Colombino, Dave Martin, Antonietta Grasso, and Luca Marchesotti

Abstract This paper considers the notion of the "semantic gap" problem – i.e. how to enable a machine to recognize the semantic properties of an image – as it is commonly formulated in the domain of content-based image retrieval. Drawing on ethnographic studies of design professionals who routinely engage in image search tasks we seek to demonstrate the means by which aesthetic and affective concepts become associated with images and elements of images within a cooperative design process of selection, discussion and refinement and how these often do not correspond to the unused semantic tags provided in image libraries. We discuss how we believe the problem of the semantic gap is misconstrued and discuss some of the technology implications of this.

Introduction

This paper is the outcome of our ongoing research on content-based image processing technologies and of a series of ethnographic studies of professionals – namely graphic designers and editorial photo researchers – whose work involves, among other things, performing image search tasks. The goal of the paper is to challenge some common assumptions about the way user behaviour and requirements are modeled in content-based image retrieval (CBIR) research, and provide some insight into how the "semantic gap" problem is really experienced by professionals who, as a routine part of their work, engage in extensive image search and retrieval activities.

In general terms, the semantic gap is characterized as the difference between a computational representation of an image's content, which is restricted to low-level

T. Colombino (✉), D. Martin, A. Grasso, and L. Marchesotti
Xerox Research Centre Europe, 6, chemin de Maupertuis, 38240 Meylan, France
e-mail: Tommaso.Colombino@xerox.com; David.Martin@xerox.com;
Antonietta.Grasso@xerox.com; Luca.Marchesotti@xerox.com

M. Lewkowicz et al. (eds.), *Proceedings of COOP 2010*, Computer Supported Cooperative Work, DOI 10.1007/978-1-84996-211-7_4,

pixel data, and the high level semantic descriptions that users might employ in any given context [4,6]. From a user's perspective the problem can be compounded by the limits of current image retrieval tools and technologies, and a more generalized difficulty in characterizing just exactly what it is that one is looking for:

> There are times and situations when we imagine what we desire, but are unable to express this desire in precise wording. Take, for instance, a desire to find the perfect portrait from a collection. Any attempt to express what makes a portrait "perfect" may end up undervaluing the beauty of imagination. In some sense, it may be easier to find such a picture by looking through the collection and making unconscious "matches" with the one drawn by imagination, than to use textual descriptions that fail to capture the very essence of perfection. One way to appreciate the importance of visual interpretation of picture content for indexing and retrieval is this [3, 5:1–5:2].

There is truth in the assertion that just what one is looking for can be difficult to articulate, in high-level language or in terms of low-level features of an image. There is also a subtle, but consequential, assumption in this type of assertion that warrants further thought – the assumption that when people search for an image, it is an act of looking for a match between an 'idea image' in their 'mind's eye' and a digital image in a collection. It is an assumption that has been borrowed somewhat non-problematically from cognitive-psychological theories, which treat visual perception and recollection as neuro-cognitive computational processes which somehow involve, literally, the presence of an image in the brain (or the mind) – presumably as a neuro-chemical state – which can be compared and matched (consciously or otherwise) to visual perceptual stimuli [8].

We are not seeking here to do a full blown conceptual critique of cognitive psychological theories of visual perception. This has been done elsewhere [1,2]. However we would urge researchers involved in content-based image retrieval technologies and in the design of novel visual asset search and management tools to remember that the metaphor of the brain as a computer and cognition as a computational process is just that – a metaphor, and the fact that it appears to fit rather nicely the formulation of the semantic gap as a computational problem that can be solved by purely computational means does not in itself, support its literal interpretation.

We want to demonstrate and argue two things. Firstly, that the notion of image matching is not borne out in real-life examples where suitable images are actually *discovered* through a search process – i.e. even when they seem to have a clear idea of what they want in advance of a search, professionals very much work out just-what-it-is that they want and how it might be appropriate through looking and associated activities. Furthermore, browsing widely is often a means of finding inspiration. Secondly, the problem of semantic labeling is not in many cases a problem about the user being unable to express what they want in aesthetic or affective terms, it is the fact that their choice of terms, as matched by a system often will not lead them to appropriate images. We want to argue that this is not due to simply needing better refinement of the mapping of semantic terms to features but is a consequence of the way semantic terms work and are applied in relation to images.

We do this because we think that it is important to clarify the conceptual relationship (and distinction) between technological–computational challenges in the field of image processing (of which content-based image retrieval is a prominent one), and the real life activities the technologies might be used to support.

We want to draw attention to the way the semantic properties of an image emerge in the course of a series of embodied and interconnected activities (an image search query, followed by browsing of the results, and the selection of a subset of relevant or candidate images, for example) which usually are in turn part of a larger, often *collaborative* applicative context (product design, or an editorial project). We will argue that in such a real-life applicative context, an image search is part of a creative process where the given requirements may be difficult to successfully map onto semantic categorisations used in image libraries.

But what is the practical consequence of this assumption? Why does it matter? There is a subtle difference between treating the semantic properties of an image as emergent (the product of embodied activities and situated collaboration) and treating images as being *statically polysemic*, i.e. being subject to several persistent semantic connotations. The notion that images are statically polysemic enforces the idea that the semantic connotations are properties of the image itself, which can be recognized (by a human or a machine) where a proper disambiguation of the right connotation can be achieved, for example, by applying the proper domain knowledge. We contend that images *are* polysemic but 'indexically' (not statically) so – i.e. the semantic meaning(s) of images are fixed repeatedly, differently, within particular sets of activities such as graphic design. The lifetime of any given meaning can be a singular situation. During a design project semantic meanings are worked up – produced, evaluated, changed, discarded, crystallized. The semantic meaning of an image is dynamic, even in the hands of a single group of users, never mind many different groups in many different places.

Furthermore, as [1] point out, "it is true that machine recognition – that is, the ability of a machine to register an object it has been programmed to pick out ('recognize') – involves matching input with electronically stored image". We suggest that when it comes to a human user, the semantic properties of an image – the ones that we're interested in – are better understood as emergent characterizations. They are not image properties which are recognized thus producing a semantic perception. However, a semantic perception or aesthetic appraisal may be mapped to image properties *after the fact*. However, this is not a necessary condition of providing a semantic description of an image.

By providing concrete examples, drawn from our ethnographic field studies, of image searches conducted by professionals (graphic designers and editorial photo researchers) we propose to show the practical ways in which these professionals make decisions about which images to select, and we hope to show how those "properties" are ascribed, made relevant, and therefore visible, accountable and agreed upon in the course of situated, and often collaborative, activities [6].

The Emergence of Semantic Properties Is Through Real-Life Image Searches

This section examines some real examples of image searches, and focuses in particular on the ways in which high level semantic properties 'become' visible in certain images. While not the only possible image retrieval scenario, here we will focus on the use of commercial image bases (such as Getty Images [8], or Shutterstock [9]) which are, generally speaking, accessed and searched directly on-line by users through a web interface which includes the search tools (which vary somewhat from case to case, but the primary tool inevitably is a textual, keyword-based, query) and a thumbnail viewing pane. Searches are initially conducted through a keyword query, the result of which can be initially constrained and subsequently refined through various parameters and browsing tools which vary from case to case.

This is the primary type of search that the professionals we observed conduct and that in CBIR scenarios is commonly describe as a "category" search. Smeulders et al. [8, p. 3] draw a distinction between a "category" search and a "targeted" search, where the main difference appears to be that in the case of a targeted search the user is either looking for a specific, known image, or has a specific image of an object in mind he or she wants to match.

The number of images returned by a search clearly depends on the specificity of the terms used in the formulation of the query and on the preponderance of images containing those elements. A search for images tagged with the keyword "beach" run on Getty Images [5] will return roughly 130,000 images, whereas a search using the keyword "duck-billed platypus" returns about 20 images (the exact number changes constantly over time as the image collections are updated). Browsing through 20 images to find the appropriate one clearly is not very time consuming, but browsing through 130,000 images (moreover where each page of thumbnails has to be loaded individually) obviously is not a very practical proposition.

In order to aid the browsing process the Getty Images interface offers a "refinement" toolbar (which can be seen on the left side in Fig. 1).

This toolbar basically consists of a refinement tree built with all the tags contained in the images in the current search space, organized by type of tag ("categories", editorial or creative; "people", which contains demographic tags; "location", which allows to select specific, named geographical locations; "keywords", which contains all types semantic tags ranging from content to aesthetic properties; "style", which contains tags pertaining to the technical properties of the image such as orientation, subject position, etc.). The numbers in parentheses next to each tag indicate how many images in the search space are associated to that tag. Clicking on a keyword in the toolbar will automatically refine the search space to all images associated with that tag. Because the refinements are concatenated, selecting three or four keywords will quickly reduce the search space. For example, in the case of the "beach" search, clicking in succession on the keywords "no people", "idyllic", and "panoramic" will reduce the search space to less than 200 images.

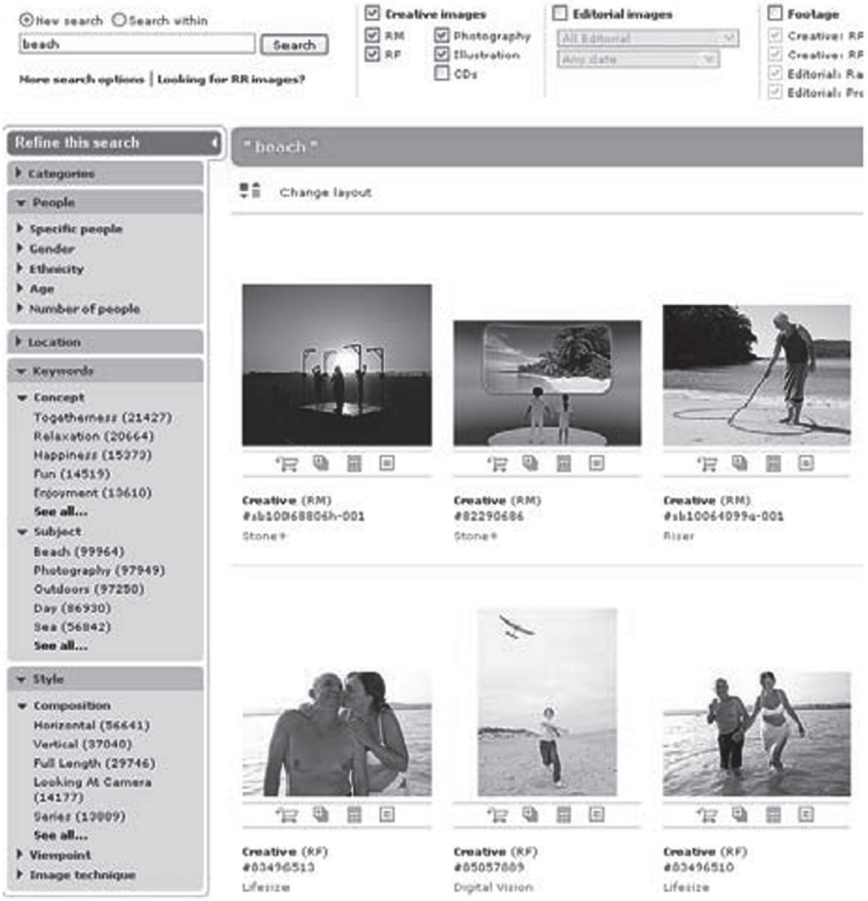

Fig. 1 Getty images search for "beach" images with keyword refinement tree on the left

It is important to note that the list of all the keywords contained in a search space of over 130,000 images is far too large to be arranged and displayed in the toolbar. There is therefore the option for the user to "see all" the tags of a particular type (Fig. 2). The tags which are actually displayed within the toolbar are simply the ones that contain the largest number of associated images (arranged in descending order).

Getty Images has, by comparison to most other commercial image bases, an advanced browsing interface which leverages the very extensive, high-level semantic tagging of all its images. The conceptual tags (such as "happiness" or "relaxation"), whether they are the product of manual or automatic indexing, are meant to capture those less tangible properties that might, for example, differentiate one picture of a beach from another which is otherwise very similar in terms of content.

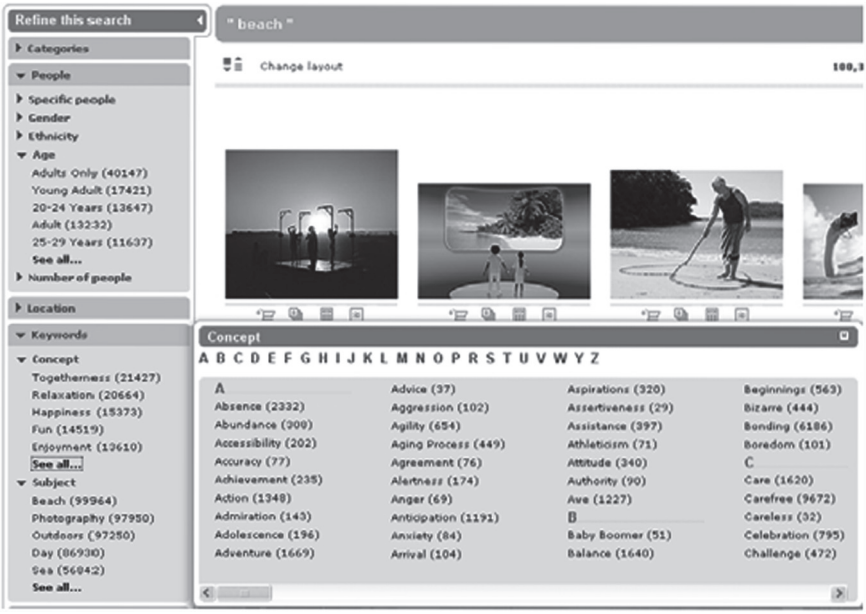

Fig. 2 Complete list of "concept" related keywords in beach search on Getty Images

The question that is relevant to our purposes here is to what extent such refinement mechanisms are useful (i.e. do they actually allow the user to efficiently navigate a search space of thousands of images to locate those images which have the desired properties).

In principle the extensive high-level semantic tagging of images provided by Getty Images is meant to address the "semantic gap".[1] The qualities they attempt to capture are precisely of those that cannot be expressed in a content-based keyword search; they apply to e.g. a mood, feeling or aesthetic in an image. However, they are not used by professionals observed by us and we would argue that this is because they immediately run up against the problem of polysemic indexicality. When we apply semantic descriptions to images there are two problems related to constancy of meaning. Firstly, there is the problem of *agreement*. Depending on the people and the context of viewing it is clear that there is variation in the semantic descriptors people deem appropriate. Yes – some archetypal image features (e.g. a smile-happiness) bring greater agreement but other many features can be shown to be far more ambiguous. Secondly, although in some cases we may find it relatively

[1]We would like to point out that we refer to Getty Images as a good example of a site that employs the principle of semantic tagging and browsing and that our argument is around how these types of technologies work. Our criticism is not of Getty Images per se – it is a popular site which enjoys good reputation among professionals for its high quality images.

easy to map semantic assessments onto concrete features of an image (as in the smile example) in other cases this is more intangible (something about the light and shadows) or very difficult to map (I don't know why but it makes me feel sad). It can be easier to articulate a semantic meaning than to specify why. Of course, it is possible to analyse images for features that recur when people have a tendency to give a certain response but cannot articulate what features it relates to *but this will only be possible in cases of high agreement amongst a user population*. This means that tags which are applied according to generic criteria by somebody other than the person doing the search may not, in fact, capture those properties that are being sought in a specific search.

As a case in point can be found in the work of a professional photo researcher who was asked to find some pictures for a feature on an increasingly popular drug called "spice", which up until a very recent ban was being legally sold in Germany as an herbal mixture (Fig. 3).

Among the pictures the journalist was looking for one had to include a person, perhaps in the process of smoking the drug. The problem the photo researcher was having was that a search for the drug itself only produced pictures of the package and the product, but not of anyone smoking it. To find pictures of people "smoking" she used the keyword "kiffer" (which loosely translates as "pot-smoker"), but felt that the resulting images were not unflattering enough in their depiction of "kiffers" for a newspaper that has a somewhat Christian conservative editorial line.

Fig. 3 Search results for "spice" retrieved from multiple providers

Fig. 4 Special feature on "spice"

A look at the picture that was ultimately printed (Fig. 4) on the paper probably shows that these judgments, while they may take up a certain amount of time and effort in the course of an image search, are difficult to characterize in terms of any properties or features the image might have that could objectively characterize it as portraying drug use in an unflattering way. In the chosen image, the fact that the whole face of the person smoking the joint was not visible seemed to reassure the photo researcher that if not unquestionably *unflattering*, the picture was at least *impersonal* enough not to contradict the overall tone of the feature itself.

The issue we want to raise with this example is whether there is in this type of search in fact a semantic gap, that could be though of as a retrieval problem, between low-level image content and the concepts such as "unflattering" or "impersonal". Ultimately, the photo-researcher was able to articulate the appropriate higher level semantic quality of the image in terms of a visual property. But this connection could not easily be articulated, or found inside the photo-researcher's head, at the moment the search itself was undertaken – it was arrived at by taking into consideration things like the content of the feature, feedback from the journalist and the art director, and the photo-researcher's knowledge through experience of selecting images that suit the paper's editorial line.

An aesthetic judgement is partly about the perception of properties (which from a practical point of view, are "looked for" when doing a search and/or choosing an image within a set) and partly about an articulation, or characterization, of those same properties which would otherwise not be "visible" and "accountable" as such in a different context. Even though it is possible to establish a link between a high-level semantic property (unflattering or impersonal) and a visual feature of the image (the partly visible face), the link would be difficult to generalize beyond the context within which it was established in the first place. And the context is not in turn reducible to a domain (such as "drug use") and they way you might model concepts such as flattering-unflattering and personal-impersonal within it.

Sketches of other observations from our studies can add to this understanding. Firstly, the notion of mind's eye matching is massively undermined by the fact that searches can often start out with apparently one target object and end up with another completely different discovered solution. For example, one of the graphic designers we observed spent some time looking for a picture of a washing machine to illustrate an article on laundry and the environment but ended up choosing a picture of clothes on a washing line because *"washing machines don't look good and clothes lines do"*!

This example can be elaborated (in terms of whether matching occurs) and enriched (in understanding emergence of semantic labels) from a look at the search and selection of fonts – a parallel but perhaps more restricted domain than photographs – for a biscuit bar wrapper design (see Fig. 5). In the picture we can see that six (three are the same) different fonts have been selected and placed on outline packets. It is a very clear example of how designers often select multiple possibilities and try them out. They don't know how an image, font, colour or whatever will work until they develop the other elements of the design, place it in context

Fig. 5 Trying out fonts for 'Bites' bars

(of images, graphics text and so forth), and they specifically assess the *candidates* with other people (colleagues, customers etc.). This process allows them to work-up and find the correct font, image, design and so forth.

Another clear feature of this comparative assessment work is that it involves making semantic articulations of what is seen and how to interpret it. Just as the photo editor could articulate why the photo 'fitted the bill' for the spice story. In the bites case the designers used the terms 'chocolatey', 'shouty', 'crunchy', 'home-made', 'it's really perfect', 'of the market' and 'for little boys' amongst others about the fonts under discussion. Are they inherent features of the fonts? Would they be seen by others? They are mapped to the fonts as part of the design articulation and assessment process. They are gestures of appreciation (perfect), sensual cross-relations (chocolatey, crunchy, shouty) and are related to the product and other expressions of what they evoke in terms of the product, consumer etc. (home-made, of the market, for little boys). Some of these can be related to concrete features of the font-images (e.g. chocolatey as a smooth edged, regular font) others not. The designers work-up an agreement on the relevant semantic labels through this process but the challenge in terms of categorisation is what number of these might travel to and be relevant in another situation of use, and also how they might relate to concrete features and as such be perceived in other font-images. We should emphasise as Wittegenstein (1966; 11) makes clear: aesthetic words can be best thought of as gestures within complex activities, such as these cooperative and organisationally circumscribed processes. Sometimes they point to concrete image features, sometimes to an overall look and feel, sometimes to an impression. And they create analogues (e.g. in different senses) and relate to all sorts of relevant features of context.

Conclusion and Technology Directions

What does this study tell us in reference to the design of search tools? It is worthwhile to reiterate our findings and suggest how they may be relevant for issues to do with design:

1. Image search is not a single order activity. Searches vary quite dramatically in terms of the concreteness or explicitness of what the person is searching for. Even with an explicit idea of required 'features', finding the right image can be time consuming. Often images can only be assessed once they are placed in various other contexts (physical layouts with other objects or cooperative and organizational situations). The 'correct' image is not a match for something held in the 'minds eye'.
2. Semantic tagging is bedeviled by the indexially polysemic nature of images; the relevance of any semantic label is always related to context (people, situation, purpose). This necessarily compromises the ability of semantic tags to be relevant cross-situationally.

3. The reliability of semantic tags – cross-situationally – will be related to at least the two following notions; *agreement across viewers* and whether they can be related to *concrete features of images* irrespective of whether these can explicitly articulated by viewers (i.e. these features could be found by machine categorisation in images people label semantically in the same way). In situations where labeling is done by a person without reference to agreement and concrete features it is likely to be problematic. We should also throw in that *fashion* dictates that once relevant categorizations may become inappropriate over time.
4. We have noted that choice of aesthetic language is not random – it is appreciative (negative-positive and otherwise), it is cross-sensual, and it makes contextual connections. These references are produced, worked-up, argued over and refined, most clearly post-selection (at least of initial candidates). The question remains as to how they can be made relevant to assist search tasks. By further investigation can we understand better whether there are patterns related to domains and how broad or narrow these domains could be. By starting small, can we create a reasonable set of semantic labels for the fonts related to biscuit bars, or images related to smokers? Should we instead concentrate on supporting archiving and the configuration of search mechanisms for small groups of designers who work together and share common perspectives and understandings, as we have seen in our studies?
5. Finally, agencies and commercial image bases may have hundreds of thousands (if not millions) of images, and any image search requires a compromise between defining and delimiting the search space through the use of explicit criteria, and browsing a sufficiently large sample of the available images to ensure that the most suitable or aesthetically appealing ones are not missed. As stated, many commercial image bases try to address this through the use of conceptual tags (such as "happiness" or "relaxation") that are meant to capture those less tangible properties that might, for example, differentiate one picture from another otherwise very similar one in terms of content. In the practices we have observed such tags were not relied upon – either in the formulation of a query or as part of a refinement/browsing process. The interesting finding – although not particularly unexpected – is that professionals inevitably conduct their searches in ways where they can have a reasonable grasp on what that search will return, hence the persistent relevance of object/feature tagging. It may not be the beach shot I want but at least it will be a beach. The key point here is that if users cannot understand the reasons for the selection of 'dreary' or 'uplifting' or 'unflattering' photos they will not use this mechanism again, it will be seen as unreliable. Users, often try to work out how something works – the decision making behind the tagging. A clear point to make, once again, is that if the means by which tags are assigned, *systematically*, can be made *visible*, in some kind of way this will clearly be more useful for users. If the tags are not added in any principled systematic manner they are unlikely to be useful.

In this paper we have sought to use ethnographic fieldwork to seek to better understand issues pertaining to the conceptualization of the nature of problems

relating to the work of image categorization and the design of image search mechanisms. Given the limits of the image search tools currently available to professionals such as graphic designers and editorial photo-researchers, there are clearly opportunities for more sophisticated retrieval technologies. However, there is also a need, as part of the research program in content-based image retrieval technologies, to bring the semantic gap problem in better focus by understanding when, how and to what end a link between visual properties and aesthetic qualities in an image is established. This should in turn help to clarify what type of image properties can usefully be leveraged by users when performing real-life image search tasks.

This is also clearly relevant for the Coop/CSCW community, not just in reiterating that it is important to understand that search and related activities have cooperative and organizational features – both within the immediate situation of search but also with a wider orientation to organizational and customer requirements. Secondly, exactly the same point can be applied to semantic labeling. Through this we hope we have contributed to better understanding for these two communities and plan to continue this research to further understand this issue and contribute more directly to technology design.

References

1. Bennet, M.R., Hacker, P.M.S.: Philosophical Foundations of Neuroscience. Blackwell, Oxford (2003)
2. Coulter, J., Parsons, E.D.: The Praxiology of Perception: Visual Orientations and Practical Actions. Inquiry, Vol. 33, pp. 251–272 (1990)
3. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys. Vol. 4, No. 2, Article 5 (2008)
4. Dorai, C., Venkatesh, S.: Bridging the Semantic Gap in Content Management Systems: Computational Media Aesthetics. COSIGN 2001, pp. 94–99 (2001)
5. Getty Images, http://www.gettyimages.com, accessed 21/09/2009
6. Martin, D., O'Neill, J., Randall, D. 'Talking about (my) Generation': Creativity, Practice, Technology, and Talk. ECSCW 2009, pp. 171–190 (2009)
7. Shutterstock, www.shutterstock.com, accessed 21/09/2009
8. Smeulders, A.W.M, Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 12, pp. 1–32 (2000)
9. Wittgenstein, L. (1966), Lectures and Conversations on Aesthetics, Psychology, and Religious Belief, ed. Cyril Barrett. Oxford: Basil Blackwell