

Towards a Transmedia Search Engine: A User Study on Perceiving Analogies in Multimedia Data

Victoria Petite¹, H. Quynh Dinh², and Ebon Fisher¹

¹Art and Technology Program, College of Arts and Letters

²Department of Computer Science
Stevens Institute of Technology

Abstract

The World Wide Web has become a reliable and fast way to archive and share multimedia data such as music, images, and videos. Most methods to search multimedia data are text-based and rely on filenames or text tags attached to the file. Those that do not rely on text are content-based. These search engines focus on exact matches and do not compare different media forms. Human cognition is much more complex, however. We typically use visual or phonetic comparisons which are then secondarily translated into language for further communication. Further, we often perceive similarities between different media forms and find analogies in non-literal, inexact matches.

The goal of our research is to develop a *Transmedia Search Engine* that suggests analogies across different media forms (*e.g.*, audio, images, videos) by looking at structural, content-based, similarity within media content. To find the most effective algorithms to achieve this goal, we are studying how people perceive similarity between and within media forms. We describe the development and results of two surveys that we have conducted on similarity between images. The first is designed to capture people's visceral reactions on how two media samples relate, while the second delves deeper into why two media samples are perceived as similar. Our results show that subject and shape are leading factors in determining similarity and are highly correlated in that study participants tend to identify both as significant factors in their perception of images as being similar.

I. Introduction

We are inspired by the ability of artists and designers to find analogies between diverse artifacts and bring them together to compose a coherent and novel narrative. An extreme form of this ability is the neurological condition known as *synaesthesia* in which two or more senses are crossed (*e.g.*, when seeing a color causes one to hear a sound). In addition to highly-regarded artists who are synaesthetic (*e.g.*, Kandinsky), there are also many examples of attempts to reproduce the effects of synaesthesia in art and entertainment (*e.g.*, the video game *Rez* [1]). Inspired by this phenomenon, we are building a *Transmedia Search Engine* to enable the exploration of analogies in mixed-media content. To find the most effective algorithms to achieve this goal, we are studying how people perceive analogies between and within media forms (audio samples, images, videos). In this paper, we describe the development and results of two surveys that we have conducted on analogies between images.

The practice of correlating different media forms has appeared throughout art history and has become even more significant in the last few decades as digital media technology has matured. An early example of analogy between different art forms is the relationship between Miro paintings and Calder sculptures. More recently, different media forms have been fused in mixed-media installations, theatre, concerts, and in visual music [29,25,16] where one media form (*e.g.*, an animation) is constructed to synchronize with another (*e.g.*, music).

The emergence of the World Wide Web as a storehouse for archiving and sharing multimedia data has enabled a different mindset – that of *gathering* (and adapting) existing media content rather than synthesizing new media content to build a coherent mixed-media narrative. To do so, efficient algorithms are needed to search multimedia data. Most search algorithms are text-based and rely on filenames or text tags attached to the file. Those that do not rely on text are *content-based* approaches that rely on meta-data extracted from the media content. There are two limitations of existing text-based and content-based search engines that our research aims to address: (1) existing search engines focus on literal and exact matches, and (2) they do not compare different media forms. Human cognition is much more complex, however. We typically use visual or phonetic comparisons which are then secondarily translated into language for further communication. Further, we often perceive similarities between different media forms and find analogies in non-literal, inexact matches.

The goal of our Transmedia Search Engine is to enable people to discover non-literal connections between text, audio samples, images, 3D geometry and videos. It is based on the psychological notion of *transderivational search* which is a fuzzy match that enables people to find contextual meaning in every stimulus and forms a primary component of human language and cognitive processing. Once built, the Transmedia Search Engine will form the core of brainstorming and discovery tools for artists to help them make mental associations in design tasks such as gathering media artifacts for a thematic installation from an archive of media samples. As artists navigate this design space, the search engine will present unexpected media possibilities. In another potential application, the search engine can be part of an interactive environment that matches the social pattern (geometry, position, and motion) of participants to media samples that are then displayed in the environment as shown in Figure 1 below.

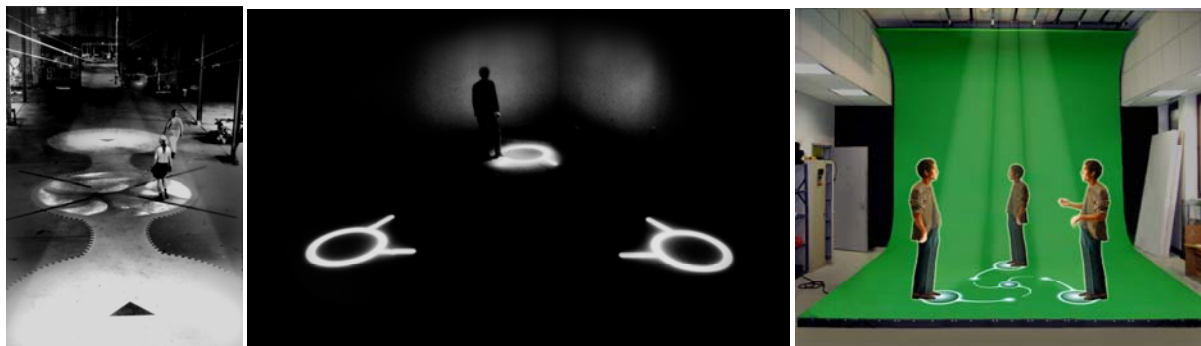


FIGURE 1 – EXAMPLES OF USING SEARCH IN MIXED-MEDIA INSTALLATIONS. Installations at the Flytrap Gallery and Test-Site Gallery in Brooklyn, NY (left, center). Mock-up of interactive installation in the green-screen room at Stevens (right).

We now review current trends in search engine technology and pattern matching algorithms that use meta-representations to find similar patterns rather than exact matches. We emphasize that this report does not present a working Transmedia Search Engine, but rather, introduces the concept of such a search engine, describes the limitations of existing search technology, and presents the results of user studies on perceiving analogies that will inform our design of algorithms toward a Transmedia Search Engine. We describe the development of the user studies in Section III and discuss the results in Section IV.

II. Related Work

The World Wide Web has become a reliable and fast way to archive and share multimedia data. Most search engines (*e.g.*, Google) are text-based and rely on filenames or text tags attached to the file to search multimedia data. Within the last 5 years, however, many content-based search algorithms have been developed that do not rely on text, and instead, compare media content using pattern-matching algorithms.

A. Content-based Search

Content-based approaches have been developed for retrieval, categorization and automated annotation of images [2,34,37,40,7,13] and video [5,12,41]. Non-textual search engines have also been developed for music [4] and 3D shapes [3,32]. These approaches focus on categorical and literal matching and do not compare different media forms. Related work that does make use of multimedia data are those that integrate multimodal sensor data (*e.g.*, audio and video) to improve tracking and surveillance [9,30,43]. Although our goals differ, the meta-data (*e.g.*, local geometric features) [28,42] these methods extract to compare media of a common form may be useful in comparing different media forms and finding non-literal associations.

B. Multimedia Clustering

Related to content-based search are methods that integrate both semantics (in the form of text) and content for organizing an image collection [7,8,11,23,39,10]. These approaches primarily deal with image databases. They attempt to learn relationships between text and image features such as the color histogram or segmentations of an image and use these relationships to perform text queries on the database and cluster images into categories.

C. Analogy-Finding

Content-based retrieval algorithms strive to identify or categorize media content given a media sample. In contrast, the goal of our Transmedia Search Engine is to find non-literal, inexact matches. Few algorithms address this goal, but one that does for images is presented in [35] where computer vision researchers Shechtman and Irani extract meta-data that captures structural similarity while being invariant to absolute appearance information such as color and texture. The resulting matches are similar while being non-literal and inexact which is what we would like to achieve across different media forms.

III. Methodology

To build a Transmedia Search Engine that suggests analogies in a content-based (rather than textual) manner, we need to build algorithms that can extract meta-data from media content and compare media samples based on the meta-data as shown in Figure 2.

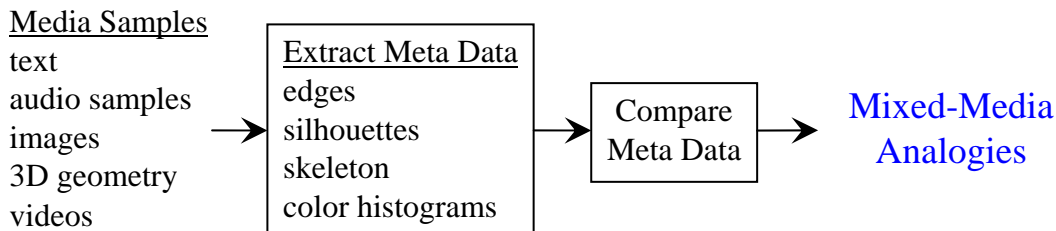


FIGURE 2 – TRANSMEDIA SEARCH PIPELINE

For images, meta-data might be in the form of edges or silhouettes extracted from the image which defines the shape of objects therein. A histogram of the color content of an image or video is another example of meta-data. Ideally, the meta-data should record information related to how analogies are perceived. To find out what meta-data should be used in our Transmedia Search Engine, we have developed two user studies on perceiving analogies, focusing on 5 different visual elements: *subject*, *shape*, *color*, *tone* (lightness/darkness), and *texture*.

A. Visual Elements

The 5 visual elements we focus on are the primary components of formal composition upon which critical analysis of artwork is based. These elements have also been identified by researchers in visual perception, neurophysiology, and computer vision as integral to the process of object recognition and identification.

Researchers in visual perception have identified two key steps in the process of visual object recognition. These are object detection and categorization [31,14,33]. Detection involves low-level visual processing such as extracting edges and segmenting objects in the foreground from the background [36]. Categorization involves high-level cognitive processing to group a detected object with existing objects in the knowledge base. The computer vision community has developed many algorithms to perform these steps for automated object recognition in images and videos. In our study, these two crucial steps map to the visual elements of subject (for categorization) and shape (for detection).

Interestingly, color and form are processed in different areas of the cortex, and studies have shown that color actually enhances recognition [17,38]. Color is also known to improve low-level vision tasks such as edge detection and object segmentation [15]. With regard texture, studies by neurophysiological researchers have shown that neurons in primate visual systems respond to texture in addition to color and shape [21,24,27]. Finally, the visual element of tone or lightness/darkness affects perceived shape and subject matter [6].


B. Online Surveys

We conducted two types of surveys. The first is designed to capture participants' visceral reactions on how two media samples relate, while the second delves deeper into why two media samples are perceived as similar. At present, our surveys allow participants to compare only images. In future work, we will conduct surveys on perceiving analogies between media of different forms such as between an image and video segment or an image and a 3D model.

The surveys are web-based and run locally on an APACHE web server. In future work, we will launch these surveys on the web for a larger user study. SQLite is used for data storage. HTML, PHP, CSS, Javascript and Flash are used to display and format the surveys.

Survey I

Recent studies have shown that detection involving low-level vision and categorization involving cognition are closely coupled and are often performed simultaneously [20, 22,19,18,26]. Although there is still debate about whether detection necessarily occurs before categorization, identification has been found to require more processing time. In the context of a search engine, identification leads to exact matches, whereas detection and categorization lead to inexact, possibly non-literal, matches. To capture the visceral perception of categories, our first survey records perceived similarity without reference to visual elements (Figure 3). The survey displays two random images and asks participants to rate the similarity on a scale of 1 (not similar) to 5 (very similar). Survey participants were asked to spend no more than 5 minutes comparing each pair of images, and the response time was 3 minutes on average. 36 participants spent approximately 1 hour each to rate an average of 1000 pairs of images.



Rate their similarity on a scale of 1-5:

1 2 3 4 5

FIGURE 3 – SURVEY I

Survey II

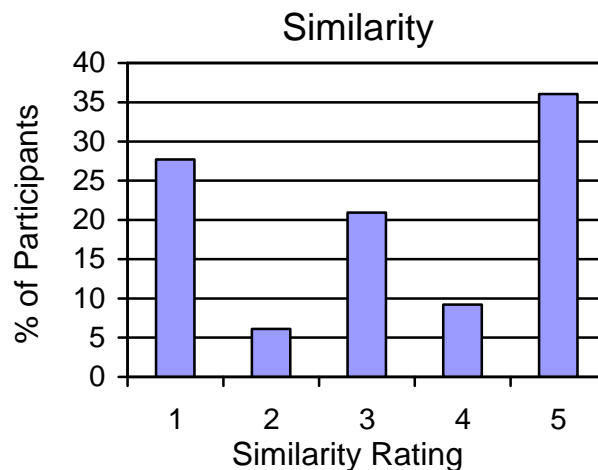
The goal of the second survey is to determine which of the 5 visual elements are most influential in perceiving analogies or similarity between images. To do so, participants are shown 6 images and asked to choose the two most similar pair. They are then asked how the 5 visual elements (subject, shape, color, texture, tone) affected their selection by rating how similar each element is in the selected pair of images on a scale of 1 (not similar) to 5 (very similar). Survey participants were asked to spend no more than 5 minutes selecting and rating each pair of images. Participants can ask for a new set of 6 images as many times as needed if they did not perceive any pair to be similar. In our study, new image sets were requested 66% of the time. 40 participants rated an average of 1000 pairs of images.

IV. Results and Discussion

We now discuss the results of our surveys and the conclusions we can infer from the data.

Survey I

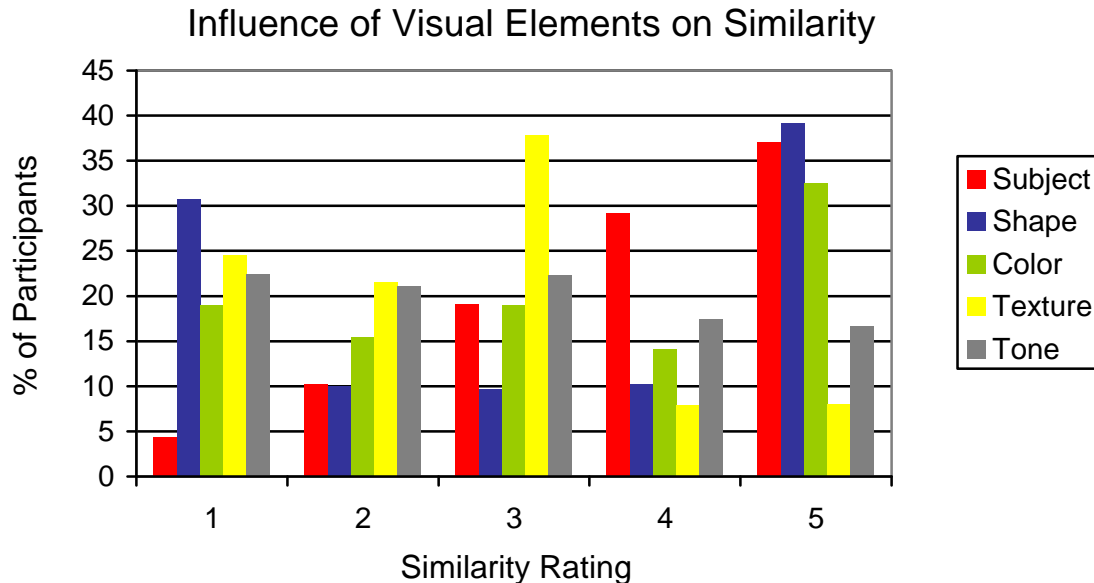
Images were found to be similar 72% of the time. The following plot shows how participants rated the similarity of a pair of random images on a scale of 1 (not similar) to 5 (very similar).



Approximately 1/3 of the responses indicated that the images are very similar (rating of 5), 1/3 indicated little or no similarity (rating of 1 and 2), and approximately 1/3 found some similarity (rating of 3 and 4), reflecting the diversity in the data. Results reveal that given two image samples, people tend to find some similarity (rating or 1 or higher) and that there may be a psychological bias towards looking for and finding similarity. Although Survey I is not particularly informative about how non-literal similarities are perceived, it does positively indicate that inexact, or analogical, matches are perceived and should be pursued. As we describe next, our second survey leads to more interesting conclusions on perceiving similarity.

Survey II

The results of our second survey are shown in the following plot. The similarity rating on a scale of 1 (not similar) to 5 (very similar) for each element is plotted against the percentage of participants who chose each rating. The ratings essentially reveal the influence of the visual element on perceiving similarity since participants rate the visual elements only once they have deemed a pair of images to be similar.



Based on the findings in the plot above, we make the following conclusions:

1. By examining the highest rating of 5, we see that subject, shape, and color are most often identified as the reason for perceiving similarity.
2. Based on the trend for each visual element, we see that subject (with the smallest percentage of “no similarity” or 1 rating), correlates most consistently with the perception of similarity.
3. The shapes in the images are either very similar or not similar at all, indicating that people may find image pairs to be similar despite having completely different shapes (most likely, subject and/or color were perceived to be similar in these cases). This result also indicates that shape is strongly perceived to be similar or different, and not often perceived to be just mildly similar.
4. Texture and tone do not appear to be highly correlated with similarity, and their ratings follow normal and nearly constant curves, respectively.

Although Survey II does not directly measure it, we make the intuitive conclusion that for more abstract image pairs, shape and color are likely to be influential, whereas for representational image pairs, subject dominates.

To further analyze the survey data, we have correlated the different visual elements. In the following table we record the percentage of time when two visual elements are identified as being very similar (rating of 5) in a pair of images. Each cell is the percentage of time that the associated row's visual element was rated as very similar given that the column's visual element was rated as very similar. For example, when subject was rated as very similar (1st column), texture was also rated as very similar only 2% of the time.

(Row & Col.) / Col.	Subject	Shape	Color	Texture	Tone
Subject		67	72	19	27
Shape	71		35	56	48
Color	63	29		22	49
Texture	2	12	5		7
Tone	12	20	25	14	

TABLE 1-CORRELATION (%) OF VISUAL ELEMENTS

The correlation percentages reveal which visual elements can be used as indicators of others. Subject is a strong indicator that shape is also similar by 71% and visa versa by 67%. This information is vital in that it enables us to infer the similarity of one visual element based on another. For example, we can infer subject similarity based on shape similarity (with 67% confidence) and rely on computer vision algorithms to robustly measure shape similarity in our Transmedia Search Engine.

V. Future Work

We have described user studies we have conducted on perceiving similarity in images. Our results show that subject and shape are leading factors in determining similarity and are highly correlated. The results will inform our design of algorithms toward an analogical search engine that we call the *Transmedia Search Engine*. Our next step is to conduct surveys using different media forms (*e.g.*, images and video segments, and images and 3D shapes) and to launch the surveys on the web for wider audience participation.

VI. Acknowledgements

This work was funded in part by NSF CreativeIT Award# IIS-0742440 and the Stevens Technogenesis Summer Scholars Program for undergraduate inter-disciplinary research.

References

- [1] <http://microscopiq.com/extras/Rez.html>
- [2] <http://www.riya.com/>
- [3] <http://shape.cs.princeton.edu/benchmark/>
- [4] <http://www.midomi.com/>
- [5] Aigrain, D., H. J. Zhang, D. Petkovic, "Content-based representation and retrieval of visual media: a state of the art review", *Multimedia Tools and Applications*, Vol.3, pp.178 – 202, 1996.
- [6] Anderson, B.L. and J. Winawer, Jonathan "Image segmentation and lightness perception", *Nature (Letters to Nature)*, Vol.434(7029), 2005. pp.79-83.
- [7] Barnard, K. and D. Forsyth. "Learning the semantics of words and pictures", *Proc. of International Conference on Computer Vision (ICCV)*, Vol.2, 2001, pp. 408 – 415.
- [8] K. Barnard, P. Duygulu, and D. A. Forsyth. "Clustering art", *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp.434–441.
- [9] Barzelay, Z. and Y. Schechner. "Harmony in motion", *Proc. of Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [10] Bekkerman, R. and J. Jeon. "Multi-modal clustering for multimedia collections", *Proc. of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp.1–8.
- [11] Cai, D., X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. "Hierarchical clustering of www image search results using visual, textual and link information", *Proceedings of the 12th International Conference on Multimedia*, 2004, pp.952–959.
- [12] Chang, S., Q. Huang, T. Huang, A. Puri, and B. Shahraray. "Multimedia search and retrieval", chapter in *Advances in Multimedia: Systems, Standards, and Networks*, A. Puri and T. Chen (eds.). New York: Marcel Dekker, 1999.
- [13] Datta, R., D. Joshi, J. Li, and J. Z. Wang. "Image retrieval: Ideas, influences, and trends of the new age," *Penn State University Technical Report CSE 06-009*, 2006.
- [14] Driver, J. and G.C. Baylis. "Edge-assignment and figure-ground segmentation in short-term visual matching", *Cognitive Psychology*, Vol.31, 1996, pp.248–306.
- [15] Fine, I., D.I.A. Macleod, and G.M. Boynton. "Surface segmentation based on the luminance and color statistics of natural scenes", *Journal of the Optical Society of America*, Vol.20, 2003, pp.1283–1291.
- [16] Fischinger, O. (director). *Oskar Fischinger: Ten Films*, 2006.

- [17] Gegenfurtner, K.R., & Rieger, J. “Sensory and cognitive contributions of color to the recognition of natural scenes”, *Current Biology*, Vol.10, 2000, pp.805–808.
- [18] Grill-Spector, K. and N. Kanwisher. “Visual recognition: As soon as you know it is there, you know what it is”, *Psychological Science*, Vol.16(2), 2005, pp.152-160.
- [19] Halgren, E., Mendola, J., Chong, C.D., & Dale, A.M. “Cortical activation to illusory shapes as measured with magnetoencephalography”, *NeuroImage*, Vol.18, 2003, pp.1001–1009.
- [20] Hochstein, S., & Ahissar, M. “View from the top: Hierarchies and reverse hierarchies in the visual system”, *Neuron*, Vol.36, 2002, pp.791–804.
- [21] Kobatake, E. and K. Tanaka. “Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex”, *Journal of Neurophysiology*, Vol.71, 1994, pp.856-867.
- [22] Liu, J., Harris, A., & Kanwisher, N. “Stages of processing in face perception: An MEG study”, *Nature Neuroscience*, Vol.5, 2002, pp.910–916
- [23] Loeff, N. , C. O. Alm, and D. A. Forsyth. “Discriminating image senses by clustering with multimodal features”, *Proceedings of COLING/ACL*, 2006, pp.547–554.
- [24] Logothetis, N., J. Pauls, H. Bulthoff, and T. Poggio. “View-dependent object recognition by monkeys”, *Current Biology*, Vol.4, 1994, pp.401-414.
- [25] Lytle, W. (director). *Animusic: A Computer Animation Video Album*, 2001.
- [26] Mack, M.L., I. Gauthier, J. Sadr, and T.J. Palmeri. “Object detection and basic-level categorization: Sometimes you know it is there before you know what it is”, *Psychonomic Bulletin & Review*, Vol.15 (1), 2008, pp. 28 – 35.
- [27] Mel, B.W. “SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition”, *Neural Computation*, Vol.9, 1997, pp. 777 – 804.
- [28] Mikolajczyk, K. and C. Schmid. “A performance evaluation of local descriptors”, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, Vol.27, pp.1615 – 1630, 2005.
- [29] Mitroo, J. B., N. Herman, and N.I. Badler. “Movies from music: Visualizing musical compositions”, *Proc. of Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1979, pp. 218–225.
- [30] O'Donovan, A., R. Duraiswami, and J. Neumann. “Microphone arrays as generalized cameras for integrated audio visual processing”, *Proc. of Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007.
- [31] Nakayama, K., Z.J. He, and S. Shimojo. “Visual surface representation: A critical link between lower-level and higher-level vision”, *An Invitation to Cognitive Science: Visual*

- Cognition* (S.M. Kosslyn and D.N. Osherson, eds.), Cambridge, MA: MIT Press, 1995, pp. 1–70.
- [32] Osada, R., T. Funkhouser, B. Chazelle and D. Dobkin. “Shape distributions”, *ACM Trans. on Graphics*, Vol.21(4), pp.807 – 832, 2002.
- [33] Palmeri, T.J. and I. Gauthier, I. “Visual object understanding”, *Nature Reviews, Neuroscience*, Vol.5, 2004, pp.291–303.
- [34] Rui, Y., T.S. Huang, and S. Chang, “Image retrieval: current techniques, promising directions and open issues”, *Journal of Visual Comm. and Image Representation*, Vol.10, 1999, pp.1–23.
- [35] Shechtman, E. and M. Irani. “Matching local self-similarities across images and videos”, *Proc. of Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007, pp.1–8.
- [36] Shin, M.C., D.B. Goldgof, and K.W. Bowyer. “Comparison of edge detector performance through use in an object recognition task”, *CVIU*, Vol. 84, 2001, pp.160–178.
- [37] Smeulders, A., M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years", *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, Vol.22(12), 2000, pp.1349–1380.
- [38] Spence, I., P. Wong, M. Rusan, and N. Rastegar. “How color enhances visual memory for natural scenes”, *Psychological Science*, Vol.17(1), 2006, pp.1–6.
- [39] Uematsu, Y., R. Kataoka, and H. Takeno. “Clustering presentation of web image retrieval results using textual information and image features”, *Proceedings of the 24th IASTED International Conference on Internet and Multimedia Systems and Applications*, 2006, pp.217–222.
- [40] Veltkamp, R. and M. Tanase. “Content-based image retrieval systems: A survey”, *Technical Report UU-CS-2000-34*, Department of Computing Science, Utrecht University, 2000.
- [41] Wang, J.Z. and N. Boujemaa (eds). *Proc. of ACM SIGMM 8th Int. Workshop on Multimedia Information Retrieval (MIR)*, 2006.
- [42] Winder, S.A.J. and M. Brown. “Learning local image descriptors”, *Proc. of Conference Computer Vision and Pattern Recognition (CVPR)*, 2007, pp.1–8.
- [43] Zhu, Z., T.S. Huang, and Y. Tian. *Proc. of IEEE Workshop on Multimodal Sentient Computing: Sensors, Algorithms and Systems (WMSC07)*, 2007.