Semantic Concept Detection for User-Generated Video Content Using a Refined Image Folksonomy

Hyun-seok Min, Sihyoung Lee, Wesley De Neve, and Yong Man Ro

Image and Video Systems Lab, Korea Advanced Institute of Science and Technology (KAIST),
Yuseong-gu, Daejeon, 305-732, Republic of Korea
{hsmin,ijiat,wesley.deneve}@kaist.ac.kr, ymro@ee.kaist.ac.kr

Abstract. The automatic detection of semantic concepts is a key technology for enabling efficient and effective video content management. Conventional techniques for semantic concept detection in video content still suffer from several interrelated issues: the semantic gap, the imbalanced data set problem, and a limited concept vocabulary size. In this paper, we propose to perform semantic concept detection for user-created video content using an image folksonomy in order to overcome the aforementioned problems. First, an image folksonomy contains a vast amount of user-contributed images. Second, a significant portion of these images has been manually annotated by users using a wide variety of tags. However, user-supplied annotations in an image folksonomy are often characterized by a high level of noise. Therefore, we also discuss a method that allows reducing the number of noisy tags in an image folksonomy. This tag refinement method makes use of tag co-occurrence statistics. To verify the effectiveness of the proposed video content annotation system, experiments were performed with user-created image and video content available on a number of social media applications. For the datasets used, video annotation with tag refinement has an average recall rate of 84% and an average precision of 75%, while video annotation without tag refinement shows an average recall rate of 78% and an average precision of 62%.

Keywords: Folksonomy, semantic concept detection, tag refinement, UCC.

1 Introduction

Nowadays, end-users can be considered both consumers and producers of multimedia content. This is for instance reflected by the significant growth in the amount of user-generated image and video content available on social media applications. An example of a highly popular website for the consumption of user-created video content is 'YouTube' [1], an online video sharing service that has nine million visitors every day, with 50,000 videos being uploaded on a daily basis [2]. Online user-created video content is currently estimated to have a size of approximately 500,000 terabytes, while it is predicted that the amount of user-created video content will continue to increase to 48 million terabytes by the end of 2011 [3]. Similar observations can be made for Flickr, a highly popular website for image sharing.

S. Boll et al. (Eds.): MMM 2010, LNCS 5916, pp. 511–521, 2010.

The availability of vast amounts of user-created image and video content requires the use of efficient and effective techniques for indexing and retrieval. However, users typically do not describe image and video content in detail [4]. This makes it difficult to bring structure in the image and video collections of users. Therefore, the automatic detection and annotation of semantic concepts can be seen as a key technology for enabling efficient and effective video content management [5-7]. Although tremendous research efforts have already been dedicated to advancing the field of automatic semantic concept detection, robust methods are not available yet.

Semantic concept detection still suffers from several major problems: the semantic gap, the imbalanced data set problem, and the use of a constrained concept vocabulary [8] [9]. Traditional approaches typically use a training database and classifiers in order to detect a limited number of semantic concepts [8-12]. These approaches make use of low-level features that are extracted from the video content and that are subsequently mapped to semantic concepts. However, people typically perceive a gap between the meaning of low-level features and the meaning of semantic concepts. As such, it is hard to model a high number of semantic concepts using a training database and classifiers [13]. The imbalanced data set problem refers to concepts that may occur infrequently. Consequently, the detection performance for rarely occurring concepts may be low due to the unavailability of a high number of training samples.

Current social media applications such as Flickr and YouTube provide users with tools to manually tag image and video content using their own vocabulary. The result of personal free tagging of image and video content for the purpose of retrieval and organization is called a folksonomy [14] [15]. The term ‘folksonomy’ is a blend of the words ‘taxonomy’ and ‘folk’, essentially referring to sets of user-created metadata.

A folksonomy typically contains a high number of images that have been manually annotated by users with a wide variety of tags. Therefore, a folksonomy may provide enough training data to learn any concept. However, tags in an image folksonomy are frequently characterized by a significant amount of noise. The presence of tag noise can be attributed to the fact that users may describe images from different perspectives (imagination, knowledge, experience) and to the fact that manual tagging is a time-consuming and cumbersome task. For example, batch tagging may cause users to annotate images with concepts not present in the image content.

This paper discusses semantic concept detection for user-created video content using an image folksonomy. Relying on the collective knowledge available in an image folksonomy is a promising approach to overcome the interrelated issues that still taunt conventional semantic concept detection. In addition, we discuss a method that allows reducing the number of noisy tags in an image folksonomy. This tag refinement method essentially uses tag co-occurrence statistics. Preliminary results show that our method is able to reduce the level of noise in an image folksonomy. Further, we demonstrate that the use of an image folksonomy allows for the effective detection of an unlimited number of concepts in user-generated video content.

The remainder of this paper is organized as follows. Section 2 presents an overall overview of the proposed system for semantic concept detection, while Section 3 describes how to refine tags in an image folksonomy. In addition, a method for computing the similarity between a video shot and images in a refined folksonomy is outlined in Section 4. Experimental results are subsequently provided in Section 5. Finally, conclusions are drawn in Section 6.

2 System Overview

Fig. 1 shows the proposed system for annotating user-created video content. Our method mainly consists of two modules: a module responsible for image folksonomy refinement and a module responsible for semantic concept detection. Given a target concept *w*, the folksonomy refinement module collects images that are representative for concept *w*. The functioning of this module will be explained in more detail in Section 3. To determine whether a concept *w* is present in a video segment, we measure the similarity between the video segment and the folksonomy images that are representative for concept *w*.

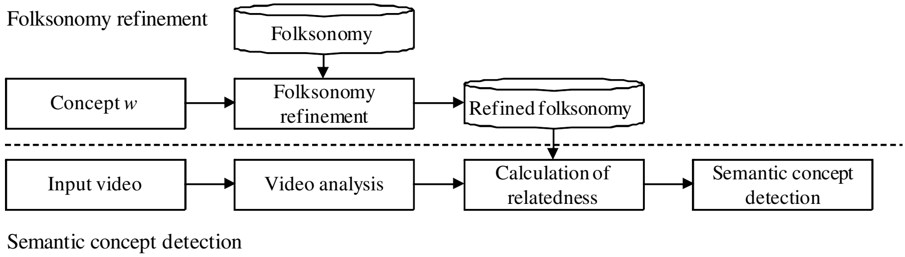


Fig. 1. Proposed system for annotating user-created video content by means of a folksonomy

3 Image Folksonomy Refinement

3.1 Definition of Correct and Incorrect Tags

It is well-known that manual tagging of image content is a time consuming and cumbersome task. The use of batch tagging may reduce the effort required by manual image tagging. However, this often results in image content annotated with incorrect tags [4]. Moreover, user-driven image tagging also introduces a personal perspective.

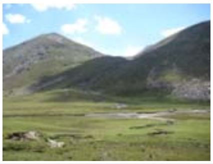
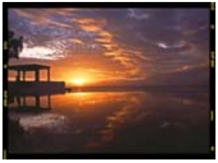
Image		
User-defined tags	village, valley, grass, horses, flowers, hills, water, lake, pond, waterfalls	church, valley, temple, tower, light, street, night, palm, tree, star

Fig. 2. Example images with user-defined tags (retrieved from Flickr). Underlined tags represent tags that are regarded as correct by human visual perception.

In our research, we assume that users assign two types of tags to image content: correct and incorrect tags. Correct tags are tags that are representative for the visual semantics of the image content, while so-called incorrect or noisy tags are not visually related to the image content. Fig. 2 shows a number of example images retrieved from Flickr, annotated with user-provided tags. Correct tags are underlined, differentiating these tags from the incorrect tags.

3.2 Measuring Confidence between a Concept and Folksonomy Images

Techniques for reducing the number of noisy tags in an image folksonomy are important in order to allow for precise search and tag-based data mining techniques [17]. To mitigate the amount of noise in an image folksonomy, we explore the relationship between different tags. In particular, the co-occurrence of several tags is taken into account. For example, the presence of the tag 'valley' typically implies that the presence of the tag 'hills' is also relevant (as shown in Fig. 2).

In general, user-generated images are created in regular situations. In other words, user-generated images rarely contain artificially organized scenes. For example, an image depicting the concepts 'sphinx' and 'water' is rare compared to an image containing the concepts 'beach' and 'water'. Consequently, we assume that the semantic concepts depicted in an image are correlated. As such, if noisy tags are present in the set of tags for a particular image, it ought to be possible to differentiate the noisy tags from the correct tags by analyzing the tag co-occurrence statistics.

We define the notion of concept confidence to represent the degree of cohesiveness between an image I and an associated semantic concept w. In this work, the confidence value is obtained by making use of the tags assigned to images. The concept confidence value is normalized, thus varying between 0 and 1. If an image I is not related to the concept w, then its confidence value will be close to 0. Likewise, if a confidence value is close to 1, then we assume that image I is highly related to concept w. The confidence value for an image I and a concept w is measured as follows:

$$Confidence(w,\mathbf{T}) = \frac{\sum_{t\in\mathbf{T}} relation(w,t)}{|\mathbf{T}|-1}, \tag{1}$$

where $\mathbf{T}$ represents the set of user-supplied tags for image I, where $|\mathbf{T}|$ represents the number of tags in $\mathbf{T}$, and where *relation()* denotes a function that maps tags on relation information. In this paper, we measure relation information between tags using tag co-occurrence statistics. In particular, the proposed method measures how often a particular tag co-occurs with another tag. This can be expressed as follows:

$$relation(w,t) := \frac{\left|\mathbf{I}^{\{w\cap t\}}\right|}{\left|\mathbf{I}^{\{t\}}\right|}, \tag{2}$$

where $\mathbf{I}^{\{w\cap t\}}$ denotes the set of images annotated with both concept w and tag t, where $\mathbf{I}^{\{t\}}$ represents the set of images annotated with tag t, and where $|\cdot|$ counts the number of elements in a set.

3.3 Refinement of Folksonomy Images

Let I be an image in the folksonomy $\mathbf{F}$ and let $\mathbf{T}$ be the set of user-defined tags associated with image I. Given the target concept w, images need to be found that are related to concept w. To find related images, the proposed refinement method uses the concept confidence computed using Eq. (1). This process can be described using the following equation:

$$\mathbf{F}_{w,refined} = \{ I \mid Confidence\,(w, \mathbf{T}) \geq threshold \}, \tag{3}$$

where w represents a concept to be detected, where $\mathbf{T}$ denotes the set of tags for image I, where *confidence()* is a function that measures concept confidence (as defined in Eq. (1)), and where *threshold* is a value that determines whether an image is related to the given target concept w or not.

4 Semantic Concept Detection

In our research, concept detection is based on measuring the similarity between a video shot and refined folksonomy images. In this section, we first describe how to extract visual features from a video shot. Next, we explain how to measure visual similarity between a video shot and a single folksonomy image. We then proceed with a discussion of how to measure the relatedness between a video shot and a set of folksonomy images all related to the same concept, enabling semantic concept detection.

4.1 Shot-Based Extraction of Visual Features

A shot in a video sequence is composed of visually similar frames. In addition, a shot is typically used as the basic unit of video content retrieval. Therefore, semantic concept detection is carried out at the level of a shot. In order to extract visual features, a video sequence $\mathbf{S}$ is first segmented into N shots such that $\mathbf{S} = \{s_1, s_2, \ldots, s_N\}$, where s_i stands for the i^{th} shot of video sequence $\mathbf{S}$ [18]. Next, low-level visual features such as color and texture information are extracted from several representative key frames. The extracted low-level visual features are described using the following MPEG-7 color and texture descriptors [20][21]: the Color Structure Descriptor (CSD), the Color Layout Descriptor (CLD), the Scalable Color Descriptor (SCD), the Homogeneous Texture Descriptor (HTD), and the Edge Histogram Descriptor (EHD). Besides color and texture information, we also extract spatial information from the key frames. This is done using the rectangle division technique described in [19]. Finally, we denote the set of visual features for shot s_i as $\mathbf{X}_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,L}\}$, where $x_{i,1}$ represents the l^{th} low-level feature extracted from shot s_i and where L is the total number of low-level features.

4.2 Similarity Measurement between a Video Shot and a Folksonomy Image

The proposed method uses a refined image folksonomy for the purpose of detecting semantic concepts in user-created video content. Specifically, the presence of a concept w in a particular video shot is estimated by making use of the visual similarity between the video shot and a set of folksonomy images that contain concept w.

Let I be an image in the refined folksonomy $\mathbf{F}_{w,refined}$. Similar to a shot in a video sequence, a set of low-level visual features is extracted for image I. This set of low-level features is represented by $\mathbf{X}_f$. The visual similarity between a video shot and the f^{th} image in the image folksonomy can be measured as follows:

$$Sim(\mathbf{X}_i, \mathbf{X}_f) = 1 - \frac{1}{L} \cdot \sum_{l=1}^{L} \sqrt{(x_{i,l} - x_{f,l})^2}, \tag{4}$$

where $x_{i,l}$ denotes the l^{th} low-level visual feature of shot s_i and where $x_{f,l}$ denotes the l^{th} low-level visual feature of the f^{th} image in the refined image folksonomy.

4.3 Semantic Concept Classification

Given a target concept w, a refined image folksonomy can be denoted as a finite set of images $\mathbf{F}_{w,refined} = \{I_1, I_2, \ldots, I_F\}$. The refined image folksonomy only contains images related to the concept w. To estimate the presence of concept w in a particular video shot, we measure the visual similarity between the video shot and all of the refined folksonomy images. The visual similarity or relatedness between a video shot s_i and the complete set of folksonomy images related to concept w is measured as follows:

$$Relatedness(\mathbf{X}_i, \mathbf{F}_{w,refined}) = \frac{1}{F} \cdot \sum_{f=1}^{F} Sim(\mathbf{X}_i, \mathbf{X}_f), \tag{5}$$

where $\mathbf{X}_i$ denotes the visual feature set of video shot s_i, where $\mathbf{X}_f$ denotes the low-level visual feature set of folksonomy image I_f, and where *Sim()* is defined in Eq. (4). If the relatedness value is higher than a pre-determined threshold, then the shot contains concept w.

5 Experiments

In this section, we first describe our experimental setup, including the construction of the image and video datasets used. We then present our experimental results.

5.1 Experimental Setup

A number of experiments were performed in order to verify the effectiveness of the proposed method for semantic concept detection in user-created video content. Our image folksonomy is constructed using the MIRFLICKR-25000 image collection [22]. This dataset consists of 25,000 images downloaded from ‘Flickr’ using its public API. Each image in the data set is annotated with tags provided by anonymous users. The average number of tags per image is 8.94. Also, the data set contains 1386 tags that have been assigned to at least 20 images.

The video annotation performance was tested for 6 target concepts: ‘street’, ‘tree’, ‘architecture’, ‘water’, ‘terrain’, and ‘sky’. Further, 70 different user-generated video sequences were retrieved from ‘YouTube’, resulting in a set of 1,015 video shots that need to be annotated. In order to evaluate the performance of the proposed semantic concept detection technique, the ground truth for all video shots was created in a manual way. In particular, three participants independently selected ground truth concepts by relying on their visual perception. Fig. 3 shows a number of extracted key frames and the corresponding ground truth concepts.

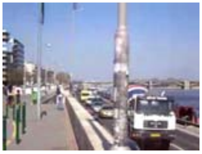
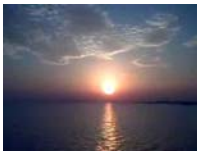
Key frames		
Ground truth	architecture, tree, sky	terrain, sky, architecture
Key frames		
Ground truth	street, water, sky	water, sky

Fig. 3. Extracted key frames and corresponding ground truth concepts

5.2 Experimental Results

We compare the accuracy of our semantic concept detection method with a technique that does not make use of a refined image folksonomy. The performance of the different concept detection methods is measured using the traditional ‘recall’ and ‘precision’ metrics. The corresponding definitions can be found below:

$$recall = \frac{N_{TP}}{N_{True}} \tag{6}$$

$$precision = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{7}$$

In the equations above, N_{TP} denotes the number of true positives, N_{FP} represents the number of false positives, and N_{True} is the number of positive samples (i.e., the total number of samples in the ground truth annotated with a particular target concept).

Given a target concept *w*, the annotation technique that does not make use of refinement uses all images in the image folksonomy that have been tagged with concept *w*. Consequently, the annotation performance is affected by folksonomy images with visual semantics that do not contain concept *w*. Fig. 4 summarizes our experimental results. Specifically, video annotation with refinement has an average recall rate of 84% and an average precision of 75%, while video annotation without refinement has an average recall rate of 78% and an average precision of 62%.

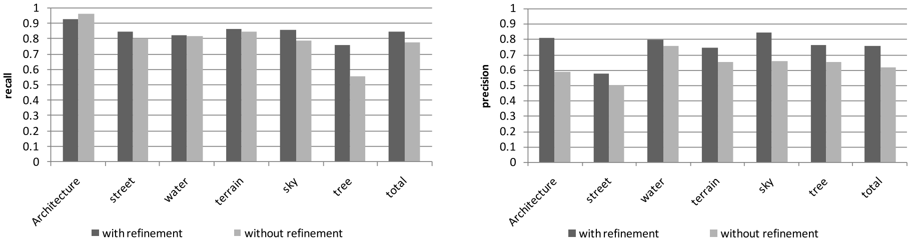


Fig. 4. Performance of semantic concept detection

Fig. 5 shows two example images that have been filtered by the proposed method. Although the user-defined tags contain the target concepts (street and tree), the content of the images does not contain these target concepts. Due to the presence of images with incorrect tags, the performance of the folksonomy-based annotation method degrades when tag refinement is not used. However, the proposed method is able to detect and remove folksonomy images with incorrect tags by making use of concept confidence values.

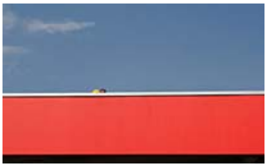
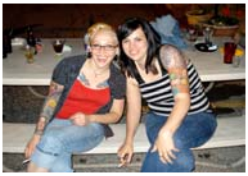
Concept	street	tree
Folksonomy		
User-defined tags	buh, brescia, colori, fdsancorastorta, ben, muibien, street	amanda, tattoo, cherries, harrypotter, tree, shamrock, girls, explored
Concept confidence value	0.1713	0.1301

Fig. 5. Example images (street and tree) filtered by the proposed method

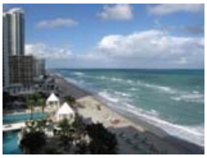
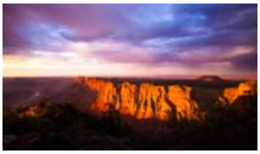
Refined folksonomy		
User-defined tags	hollywood, florida, beach, resort, roadtrip, sea, clouds	grand canyon, sunset, sun, clouds, hill, landscape, blue, peaceful
Concept confidence value	0.4339	0.4756

Fig. 6. Example images that are recommended for the sky by the proposed method

Further, even if the user-defined tags do not contain the target concept, the actual content of the images may still include a particular target concept. The proposed folksonomy refinement process is able to find such images by exploring the tag relations. Fig. 6 shows two example images that are recommended for the concept ‘sky’. While the corresponding tag sets of the two images do not include the concept ‘sky’, tags that are highly related to the concept ‘sky’ were used to annotate the images. Highly related tags include ‘clouds’ and ‘beach’ for the image to the left, and ‘blue’ and ‘sunset’ for the image to the right.

However, while the video annotation method with refinement mitigates the impact of noise in an image folksonomy, we have observed that the performance of our method is low for concepts such as ‘tree’ and ‘street’. Indeed, although the images in the refined image folksonomy are all related to the target concepts ‘tree’ and ‘street’, the diversity of the respective image sets is high in terms of visual similarity and tags used. This observation is for instance illustrated in Fig. 7 for the concept ‘tree’.

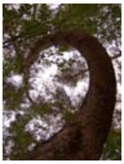
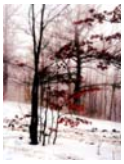
Refined folksonomy			
User-defined tags	tree, cliff, windswept, wales	tree, pine, spiral	Snow, albero, foglie, tree
Concept confidence value	0.5108	0.5000	0.5410

Fig. 7. Example images and corresponding confidence values for the concept ‘tree’

6 Conclusions and Future Work

This paper discussed a new method for semantic concept detection in user-created video content, making use of a refined image folksonomy. Tag refinement mitigates the impact of tag noise on the annotation performance thanks to the use of tag co-occurrence statistics. That way, the proposed annotation method can make use of folksonomy images that are better related to the concept to be detected. Our experimental results show that the proposed method is able to successfully detect various semantic concepts in user-created video content using folksonomy images.

Despite the use of tag refinement for the purpose of tag noise reduction, the refined image folksonomy may still contain a diverse set of images that are all related to the same target concept. Future research will conduct more extensive experiments in order to study the aforementioned observation in more detail. In addition, future research will focus on improving the concept detection accuracy by also making use of a video folksonomy.

References

1. YouTube, http://www.youtube.com/
2. 7 things you should know about YouTube (2006), http://www.educause.edu/ELI/7ThingsYouShouldKnowAboutYouTu/156821
3. Ireland, G., Ward, L.: Transcoding Internet and Mobile Video: Solutions for the Long Tail. In: IDC (2007)
4. Ames, M., Naaman, M.: Why We Tag: Motivations for Annotation in Mobile and Online Media. In: ACM CHI 2007, pp. 971–980 (2007)
5. Wang, M., Hua, X.-S., Hong, R., Tang, J., Qi, G.-J., Song, Y.: Unified Video Annotation via Multi-Graph Learning. IEEE Trans. on Circuits and Systems for Video Technology 19(5) (2009)
6. Wang, M., Xian-Sheng, H., Tang, J., Richang, H.: Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation. IEEE Trans. on Multimedia 11(3) (2009)
7. Yang, J., Hauptmann, A., Yan, R.: Cross-Domain Video Concept Detection Using Adaptive SVMs. In: Proceedings of ACM Multimedia, pp. 188–197 (2007)
8. Chen, M., Chen, S., Shyu, M., Wickramaratna, K.: Semantic event detection via multimodal data mining. IEEE Signal Processing Magazine, Special Issue on Semantic Retrieval of Multimedia 23(2), 38–46 (2006)
9. Xie, Z., Shyu, M., Chen, S.: Video Event Detection with Combined Distance-based and Rule-based Data Mining Techniques. In: IEEE International Conference on Multimedia & Expo. 2007, pp. 2026–2029 (2007)
10. Jin, S.H., Ro, Y.M.: Video Event Filtering in Consumer Domain. IEEE Trans. on Broadcasting 53(4), 755–762 (2007)
11. Bae, T.M., Kim, C.S., Jin, S.H., Kim, K.H., Ro, Y.M.: Semantic event detection in structured video using hybrid HMM/SVM. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 113–122. Springer, Heidelberg (2005)
12. Wang, F., Jiang, Y., Ngo, C.: Video Event Detection Using Motion Relativity and Visual Relatedness. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 239–248 (2008)
13. Jain, M., Vempati, S., Pulla, C., Jawahar, C.V.: Example Based Video Filters. In: ACM International Conference on Image and Video Retrieval (2009)
14. Ramakrishnan, R., Tomkins, A.: Toward a People Web. IEEE Computer 40(8), 63–72 (2007)
15. Al-Khalifa, H.S., Davis, H.C.: Measuring the Semantic Value of Folksonomies. Innovations in Information Technology, 1–5 (2006)
16. Lu, Y., Tian, Q., Zhang, L., Ma, W.: What Are the High-Level Concepts with Small Semantic Gaps? In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2008)
17. Xirong, L., Snoek, C.G.M., Worring, M.: Learning Tag Relevance by Neighbor Voting for Social Image Retrieval. In: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval, pp. 180–187 (2007)

18. Min, H., Jin, S.H., Lee, Y.B., Ro, Y.M.: Contents Authoring System for Efficient Consumption on Portable Multimedia Device. In: Proceedings of SPIE Electron. Imag. Internet Imag. (2008)
19. Yang, S., Kim, S.K., Ro, Y.M.: Semantic Home Photo Categorization. IEEE Trans. on Circuits and Systems for Video Technology 17(3), 324–335 (2007)
20. Ro, Y.M., Kang, H.K.: Hierarchical rotational invariant similarity measurement for MPEG-7 homogeneous texture descriptor. Electron. Lett. 36(15), 1268–1270 (2000)
21. Manjunath, B.S., et al.: Introduction to MPEG-7. Wiley, New York (2002)
22. Huiskes, M.J., Lew, M.S.: The MIR Flickr Retrieval Evaluation. In: ACM International Conference on Multimedia Information Retrieval (MIR 2008), Vancouver, Canada (2008)