Web Image Annotation based on the Decision Rules Inferred by the Statistical Analysis of Web Pages

Joohyoun Park*, Giseok Choe*, Jongwon Lee^o, Jongho Nang*
Dept. of Computer Science and Engineering, Sogang University
1 ShinsuDong, MapoGu, Seoul, Korea*
{parkjh, brix, jhnang}@sogang.ac.kr*
Dept. of e-Sports Game, Chungkang College of Cultural Industries,
33 Haewol-ri, Majang-myun, Ichon, Kyungki-do, Korea^o
jongwon@ck.ac.kr^o

Abstract

This paper proposes a rule based web image annotation method which improves the precision and recall of annotation by the use of decision tree. This decision tree learns the relationship between images and their annotations based on the proposed 17 attributes that specify the structural relationship between them in HTML documents and the visual characteristics of the images. By converting and pruning this learned tree, a set of rules with high estimated accuracy which determines whether or not a word can be the keyword of an image can be generated. Upon experimental results, the proposed method made 57 rules and the precision and recall of annotation by these rules were about 88% and 95% for the various concepts, respectively. We argue the contribution of this work in two aspects. First, we suggest the clear criteria for precise annotation inferred by the statistical analysis of many web pages. Second, to cope with the deterioration of recall caused by the lack of measure for the visual characteristics, the visual similarity between an image and its concept combines to the attributes that used for tree learning.

1 Introduction

Advent of new technologies in WWW and personal imaging devices such as digital camera and mobile phone lead to increase the number of images on the WWW. Consequently, the needs of efficient mining, managing, and searching methods for these web images have been increased as well. Some major search engines such as Google and Yahoo already provide the text based image retrieval(TBIR) service which finds relevant images to a given textual query. In TBIR, the quality of retrieval results quite depends on the precision of web image annotation.

There were some researches[12][14][7][2][15] which describe the problem of image auto-annotation as a supervised or an unsupervised learning problem which builds up the relationship between visual features and concepts. However, the annotations which generated by this approach would not describe the image content accurately because of "Semantic Gap" problem. In case of images on WWW, a different approach is available because web pages embedding images also contain descriptive texts staying close to the images. In this manner, web images are successfully collected and annotated by many previous works[13][3][8][11][17][4][1].

In these works, all words in the HTML document embedding an image can be the candidate keywords of the image. Of course, since all words in the HTML document may not be evenly relevant to the image, the criterion of proximity to the image should be evaluated. For example, the words closer to the image or appearing with some specific tags may give higher weight as compared to other words. However, some words with higher weights may not be relevant to the images in some HTML documents with specific layout and vice versa. To solve this problem, this weighting scheme should adapt the layout of HTML document. That is, a set of annotation rules considering the layout of HTML document is required for the precise annotation.

This paper proposes a decision rule based web image annotation method using both textual and visual information. Our basic idea is to annotate images based on a set of rules which is inferred by the statistical analysis of the relationship between images and their annotations in HTML documents. To represent this relationship, we suggest 17 attributes, which indicate whether or not a term is appearing with specific tag, how far is a term from an image, and the properties of entire document such as the term frequency or

0-7695-2983-6/07 $25.00 © 2007 IEEE
DOI 10.1109/CIT.2007.123

IEEE computer society

the number of image. One of the attributes represents the visual similarity between an image and its concept. It can be used to cope with the lack of measure which evaluates the degree of relevance between the image and its term. Finally, images are mined with their precise annotations on WWW based on the rules generated by decision tree with C4.5[10] algorithm. Upon experimental results, the recall and precision of annotation by the proposed method shows about 95% and 88%, respectively.

2 Construction of Image Dictionary

To evaluate the degree of visual match between an image and its concept, we use the visual keyword which is the visual data types abstracted and extracted from visual documents in a content domain[16]. In other words, the visual keyword is the data structure which represents the relationship between the visual information extracted from images and their concepts. Image dictionary can be defined as a set of the visual keywords for various concepts. In this paper, this image dictionary is built up by the following learning process, which is similar to [7].

To learn the concepts associated with the visual information, a large number of labeled images have to be prepared. Commercial image collections or a set of collected images from WWW by [9] can be used as the training image set. Once a set of the training image for various concepts is prepared, some features which can well describe the visual characteristics of each concept have to be extracted. However, since most natural images describe multiple concepts and various backgrounds, features extracted from an entire image may not represent the visual characteristics of a concept purely. To remove the noises which were incurred by complicated images with multi-objects, each sample image is segmented into 3x3 uniform blocks. Among these 9 blocks, the center block is picked up as the representative region of the image to focus a main object.

As the training samples to learn the relationship between the concept and its visual information, 4 MPEG-7 visual descriptors[5] such as *Dominant Color*, *Color Layout*, *Edge Histogram*, and *Color Structure* are extracted from these representative regions. Based on these features, each representative region is categorized by *k-means clustering* algorithm with equal weights. Then each category has the regions with similar visual properties and with the words annotated manually at the image preparation step. Finally, for each category $C_i (1 \leq i \leq k, k$ is the number of clusters), a set of representative keywords W_{C_i} can be built based on the frequency of the words annotated to the regions in the category and the degree of relevance to each keyword $R_c(C_i)$ can be evaluated as follows;

$$R_c(C_i) = \frac{M_c(C_i)}{\sum_{c \in W_{C_i}} M_c(C_i)} \quad (1)$$

Note that c is one of representative keywords and $M_c(C_i)$ is the term frequency of c in the cluster C_i.

3 Decision Rule based Annotation

3.1 Attributes of HTML Documents Embedding Images

HTML documents embedding images usually have many hints which imply the degree of relevance between the images and some words co-occurred with them. These hints can be taken by analyzing the layout of HTML documents and broadly classified into 4 categories.

First, The words appearing with *src*, *alt* fields of <IMG> tag have higher importance than other words. Especially, since *alt* field was originally planned as an alternative for browsers that did not show images, the words appearing with this field may be the best clue to represent the semantics of the source image. Second, the words emphasized by some specific tags such as <Hx> <B>, <BIG>, <EM>, and <I> may be more relevant to nearby images. That is, since the words appearing with these tags usually represent the content of paragraphs, they may have high relevance to nearby images. Third, the structure information of HTML documents should be considered. The structural analysis of HTML documents is made based on DOM (Document Object Model)[1]tree, which defines the logical structure of documents. From this DOM tree, we can easily comprehend the relationship between an image block and its surrounding text blocks. Finally, the properties of the entire document should be considered. Intuitively, the image mining at well structured HTML documents which contain many content images may produce good results. Also, the title or the URL of documents may be a good attribute to get the relevant images to the given concept.

From these considerations, we propose 17 attributes to evaluate the degree of relevance between an word c and an image o which are embedded in a HTML document as shown in Table 1. Attributes from A1 to A6 indicate whether or not c is appeared with some special tags such as *alt*, *src* fileds if the *img* tag, *H*x, and *title*. Attributes from A7 to A9 represent the distance between o and c in a number of ways. A10 represents the size of the DOM block including c and it can be the clue that tells the relative importance of c in this text block with the co-use of the term frequency. Attributes from A11 to A14 represent the information directly extracted from the image o. They can be the terms that filter out unsuitable images to be collected for CBMR such as icon images for navigation or banner images for advertising. A15 indicates the term frequency of c in the HTML document. Although *tf/idf* is a well known

[1] http://www.w3c.org/DOM

technique for estimating the importance of keywords in a document, we use the pure term frequency. The reason for that is that *idf* may decrease the importance of c in some sites which consist of many HTML documents describing the same concept. A16 represents the number of content images, which simply can give a hint of which attribute plays a key role in annotation process. For example, if a HTML document includes only one image, the word in *title* tag is likely to directly describe the image. Otherwise, the closeness from the image may be more important than whether or not the word appears with *title* tag.

A17 represents the visual distance $d(c,o)$ between the image o and its concept c, which can be evaluated by the use of visual keywords. That is, it can be calculated as the minimum distance from the image o to the cluster including c as its keyword;

$$\begin{aligned} d(c,o) = min\{d(o,C_i)|C_i &\text{ is the cluster which has } c \\ &\text{as the keyword}\} \\ &, where\ d(o,C_i) = R_c(C_i) \times d_v(o,C_i) \quad (2) \end{aligned}$$

Note that $d_v(o,C_i)$ means the visual distance between o and C_i which is calculated by the weighted sum of the distances between MPEG-7 visual descriptors of the centroid of C_i and the image o.

3.2 Decision Rule Induction and Annotation

Basically, all words in a HTML document can be the candidate keywords of the images in this document but all words in this document can not be the final keywords of the images. That is, some words with low relevance to the images must be filtered out to improve the precision of annotation. For this filtering process, a set of rules which determines whether or not to pick up the candidate words as the final keywords of the image should be built. In this paper, these rules are built based on the proposed 17 attributes of relationship between an image and an word embedded in a HTML document (see **Table 1**). To build these rules effectively, we choose the decision tree with C4.5 algorithm[10] as the inference tool which is one of the most widely used for inductive inference. The decision tree is inferred from the training set then it is converted into an equivalent set of rules. To prevent overfitting, each rule is pruned to improve its estimated accuracy. Based on these rules, a word in the HTML document can be included in or excluded from a set of keywords of an image in the same document.

Table 1. 17 attributes of relationship between an image o and an word c which are embedded in a HTML document

ID	Attribute	Possible Values
A1	$c \in \Psi_{ALT}$	{yes, no}
A2	$c \in \Psi_{Hx}$	{yes, no}
A3	$c \in \Psi_{IMG}$	{yes, no}
A4	$c \in \Psi_{TITLE}$	{yes, no}
A5	$c \in \Psi_{URL}$	{yes, no}
A6	$c \in \Psi_B \bigcup ..., \Psi_{STRONG}$	{yes, no}
A7	$B(o) \in P(B(c))$	{yes, no}
A8	$f_l(B(c),B(o))$	continuous
A9	$f_p(B(c),B(o))$	continuous
A10	$f_s(B(c))$	continuous
A11	$f_s(B(o))$	continuous
A12	Position of o	{center, outer edge}
A13	File Type of o	{jpeg, other type}
A14	Aspect Ratio of o	continuous
A15	Frequency of c	continuous
A16	Number of Content Images	continuous
A17	d(c, o)	continuous

Notations
o : an image embedded in the HTML document
c : an word embedded in the HTML document
Ψ_{Tag} : a set of noun words appearing with *Tag*
$B(\cdot)$: the DOM block including an word or an image
$P(\cdot)$: the parent of a given DOM block
$C(\cdot)$: the child of a given DOM block
$f_l(\cdot,\cdot)$: the function which yields the number of edges between two DOM blocks
$f_p(\cdot,\cdot)$: the function which yields the distance between two center points of each block by pixel
$f_s(\cdot)$: the function which yields the size of DOM block
$d(c,o)$: the visual distance from o to the visual keyword of c

To illustrate this rule based annotation scheme formally, suppose Θ is a HTML document embedding n images and m words. And o_i $(1 \leq i \leq n)$ and c_j $(1 \leq j \leq m)$ denote the i^{th} image and the j^{th} word in Θ, respectively. From Θ, $n \times m$ instances can be extracted for training and testing and they are denoted by $< o_i, c_j >$ $(1 \leq i \leq n, 1 \leq j \leq m)$, which are represented by the 17 attributes such as <A1, A2, A3, ..., A17>. For example, assuming that Θ has a "tiger" image (320×240) and the word "tiger" is appeared with *alt* field, then an instance of < "tiger" image, "tiger"> which is represented by <yes, no, no, no, no, no, yes, 0, 0, 0, 76800, center, jpeg, 1, 1, 1, 0.2> can be generated from Θ. Moreover, some words which are not embedded in Θ can be added to this set of instances by matching the visual features extracted from the image with the visual keywords in the image dictionary. If l words are added by this process for each image in Θ, the total number of instances can be extracted from Θ is $n \times (m+l)$. Based on these attributes of $< o_i, c_j >$, the value of the target function $\tau(o_i,c_j)$ is determined to either *"Select"* or *"Discard"*. If $< o_i, c_j >$ is

Table 2. 6 seed sites for the experiments

Seed site
http://www.junglewalk.com
http://nationalzoo.si.edu
http://www.freefoto.com
http://www.hickerphoto.com
http://www.amusetoi.com
http://www.indianwildlifeportal.com

classified into *Select*, c_j can be the annotation of o_i, otherwise not. For example, τ(”tiger” image,”tiger”) = *“Select”* is expected.

Figure 1 shows an algorithm to annotate the images extracted from a HTML document. The procedure **RulebasedAnnotation** has two arguments, a set of rules r for auto-annotation generated by decision tree and a web page Θ collected by web spider. In this procedure, a set of images and words denoted by O and C are built by parsing the HTML code of Θ. For each image o_i in O, $n(C)$ (the function of $n(\cdot)$ yields the number of elements in a set) pairs of $< o_i, c_j >$ $(1 \leq j \leq n(C))$ should be evaluated by the procedure **EstimateTarget** because all words in C can be the candidate keywords. Procedure **EstimateTarget** predicts whether or not c_j can be the keyword of o_i based on 17 attributes extracted by the procedure **Get17Attributes**. If the decision for $< o_i, c_j >$ is *“Select”*, c_j is to be the keyword of o_i.

4 Experimental Results and Analysis

4.1 Experimental Enviroments

To evaluate the performance of the proposed annotation method on the web environments, we gathered 15,185 images from various web pages by the web crawling robot. This robot visited enormous number of web pages by BFS(Breadth First Search) traversal method from 6 seed sites as shown in **Table 2**. To make the ground truth for evaluation of the performance such as precision and recall, we selected the 9497 images that are paired with the 19 dominant concepts on the collected images and annotated them manually as shown in **Table 3**. Finally, 3,842 positive instances and 5,655 negative instances were used for training and testing to evaluate performance of the proposed method. All experimental results were evaluated based on this ground truth with the 10-fold cross validation. We used MSHTML library (http://msdn2.microsoft.com/en-us/library/ms905080.aspx) for DOM analysis and MPEG-7 XM[6] for extracting the visual features. For decision tree, the source code of C4.5(Release8) was used.

```
r : a set of annotation rules inferred from a large number of training set
$C_i$ : the $i^{th}$ category in the image dictionary
$\Theta$ : an HTML document
ExtractImage($\Theta$) : a function that extracts the images in $\Theta$
ExtractWords($\Theta$) : a function that extracts the noun words in $\Theta$
Get17Attributes($o_i$, $c_j$, $\Theta$) : a function that extracts the proposed 17 attributes of $o_i$ and $c_j$ in $\Theta$
EstimateTarget($r$, $Attr$) : a function that estimates the target value (“Select” or “Discard”) from $Attr$ based on $r$
GetKeywordsByVS(Image $o$, $l$)
$d_{min}$ = MAX_VISUAL_DISTANCE;
for all $C_i \in$ image dictionary such that $1 \leq i \leq k$ do
  $d = d_v(o, C_i)$;
  if $d < d_{min}$ then
    $C_{min}$=$C_i$; $d_{min} = d$;
  end if
end for
sort the keywords of $C_{min}$ by $R_c(C_{min})$
return top $l$ keywords of $C_{min}$; RulebasedAnnotation(Rules r, HTML $\Theta$)
$O$ = ExtractImages($\Theta$);
$C$ = ExtractWords($\Theta$);
for all $o_i \in O$ such that $1 \leq i \leq n(O)$ do
  $C = C \cup$ GetKeywordsByVS($o_i$, $l$);
  for all $c_j \in C$ such that $1 \leq j \leq n(C)$ do
    $Attr$ = Get17Attributes($o_i$, $c_j$, $\Theta$);
    $Decision$ = EstimateTarget($r$, $Attr$);
    if $Decision$ == ”Select” then
      select $c_j$ as the keyword of $o_i$;
    else
      discard this pair;
    end if
  end for
end for
```

Figure 1. Algorithm for rule based annotation

4.2 Annotation Results and Analysis

Even though the generated rules were slightly different according to what folds were used for training, we could pick up some common rules with high estimated accuracy. Among these rules, two important common rules are shown in **Figure 2**. By rule 4, a word can not be the keyword of the paired image if the word is not appearing with *alt* field, *img* tag and *title* tag. On the other hand, if an word is appearing with *title* tag and the visual distance between the word and

Table 3. The number of the collected images for each concept by web spider (denoted by A) and the number of the annotated images to this concept manually (denoted by B)

concept	A	B	concept	A	B
cat	4240	126	bird	435	296
bear	664	511	beach	173	150
aurora	278	255	tree	248	151
tiger	343	190	sunset	373	335
shark	206	128	lion	304	193
lake	256	161	italy	200	120
horse	263	171	giraffe	92	85
france	336	175	flower	108	102
elephant	206	127	dolphin	474	370
dog	298	196	Total	9497	3842

the paired image is less than 0.248, the word is likely to be the keyword of the image by rule 31.

```
■ Rule 4:
if A1=0 & A3=0 & A4=0 then
  "Discard"
end if

■ Rule 31:
if A4=1 & A17≤0.248 then
  "Select"
end if
```

Figure 2. An example of rules

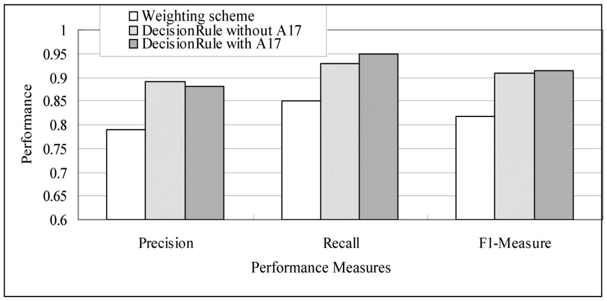


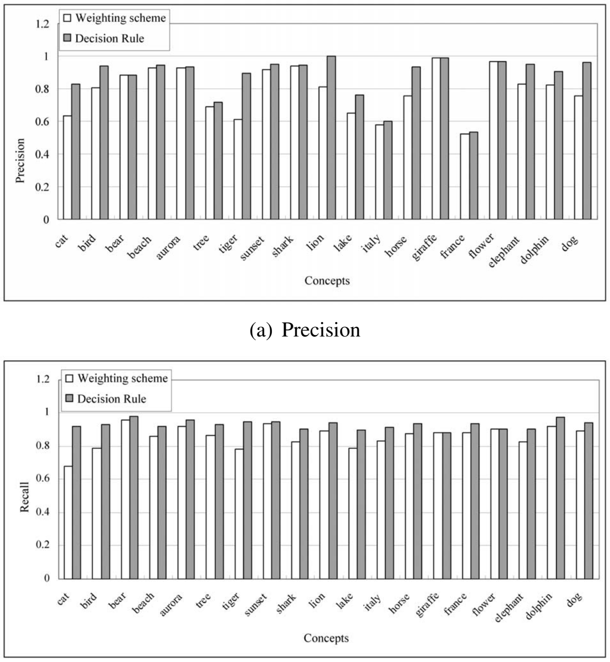
Figure 3. Overall performance comparison

Figure 3 shows the comparison of overall performance between the annotation based traditional weighting scheme and the proposed annotation method. This weighting scheme is basically referred to [13] and an word whose normalized weights is larger than 0.5 is selected as the keyword of the paired image. In case of the proposed method, we made two kinds of experiments with A17 and without A17 to show how A17 affects the performance of annotation. As shown in **Figure 3**, both the precision and the recall of the proposed method are about 10% higher than the method based on the weighting scheme averagely regardless of A17. These results imply that some simple rules which are made by an analysis of generic HTML documents like weighting scheme are not enough to annotate web images accurately. That is, some complicated rules which reflect the properties of various kinds of HTML documents are required for accurate annotation and the proposed method gives satisfaction to this requirement. Also, **Figure 3** shows that A17 makes slight differences in the precision and recall of the proposed method. By the use of A17, the number of candidate keywords can increase (see 4-5 lines of **RulebasedAnnotation** procedure in **Figure 1**) and the visual characteristics of a word which is very similar to that of the paired image visually can be selected as the keywords regardless of other attributes. It leads that the recall of annotation with A17 is slightly higher than that without A17. However, "Semantic Gap" problem may put down the precision of annotation with A17.

Figure 4 shows that the precision and recall of annotations by the proposed method with A17 and weighting scheme are represented per each concept. For all concepts, both the precision and the recall of annotation by the proposed method were higher than those of annotation based on weighting scheme. It implies that the proposed method is superior to the other method regardless of kinds of concept. However, for some concepts such as "tree", "lake", "italy", and "france", both methods did not show good precisions compared to other concepts as shown in **Figure 4**-(a). The reason for these results is that these concepts may be not occurred as the main objects in pictures even though they appear in HTML documents such as "lion under tree", or "cosmetics made in france". Also, since the visual characteristics of images for these concepts have much variations, the visual keywords for these concepts are hard to learn the common visual information for each concept. This makes it much harder to improve the precision for these concepts. In case of recall, the proposed method gives a high and stable performance for all concepts as shown in **Figure 4**-(b).

5 Conclusion

In this paper, we presented a web image annotation method based on the decision rules which were inferred by the statistical analysis of many web pages. To analyze the relationship between images and their annotations in various kinds of web pages effectively, we suggested 17 attributes that specify the structural information extracted from HTML documents and the visual characteristics of the



(a) Precision

(b) Recall

Figure 4. Experimental comparison of the precision and the recall per concept

images. Decision tree with C4.5 was used as the analysis tool and made a set of rules as clear criteria for precise annotation. Based on these rules, the performance of annotation was evaluated for 9,497 testing samples collected from WWW. Upon experimental results, the average precision and recall of annotation by the proposed method were about 88% and 95%, respectively. Even though these improved results showed the superiority of the proposed method averagely, the precisions of annotation for some concepts which were used to occurred as side objects in pictures were still low. We plan to solve this problem in our future work. Nonetheless, since the average performance of the proposed method is high, it can be used to improve the quality of search for text based image retrieval systems effectively.

References

[1] D. Cai, X. He, W. Ma, J. Wen, and H. Zhang. Organizing www images based on the analysis of page layout and web link structure. In *Proceedings of IEEE International Conference on Multimedia Expo*, pages 113–116, 2004.

[2] C. Cusano, G. Ciiocca, and R. Scettini. Image annotation using svm. In *Proceedings of Internet Imaging*, pages 330–338, 2004.

[3] C. Frankel, M. Swain, and V. Athitsos. *WebSeer: An Image Search Engine for the World Wide Web, Technical Report 96-14*. University of Chicago Computer Science Department, 1996.

[4] Y. Lai, S. Liu, L. Tien, and S. Chan. Semantic knowledge building for image database by analyzing web page contents. In *Proceedings of IEEE International Conference on Multimedia Expo*, pages 1282–1285, 2005.

[5] ISO/IEC JTC1/SC29/WG11. *Information Technology Multimedia Content Description Interface-Part3: Visual*. 2001.

[6] ISO/IEC JTC1/SC29/WG11. *MPEG-7 Visual part of eXperience Model Version 11.0*. 2001.

[7] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[8] S. Mukerjea and J. Cho. Automatically determining semantics for world wide web multimedia information retrieval. *Journal of Visual Languages and Computing*, 10:585–606, 1999.

[9] J. Park and J. Nang. A novel approach to collect training images from www for image thesaurus building. In *Proceedings of IEEE Symposium on Computational Intelligence in Image and Signal Processing*, 2007.

[10] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[11] N. Rowe and B. Frew. Automatic caption localization for photographs on world wide web pages. *Information Processing and Management*, 34(1), 1998.

[12] S. Rui, W. Jin, and T. Shua. A novel approach to auto image annotation based on pair-wise constrained clustering and semi-naive bayesian model. In *Proceedings of IEEE 11th International Conference on Multimedia Modeling*, pages 322–327, 2005.

[13] S. Sclaroff, L. Taycher, and M. Cascia. ImageRover: A content-based image browser for the World Wide Web. In *Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 2–9, 1997.

[14] J. Wang and J. Li. Learning-based linguistic indexing of pictures with 2-d mhmms. In *Proceedings of ACM Multimedia 2002*, pages 436–445, 2002.

[15] L. Wang, L. Liu, and L. Khan. Automatic image annotation and retrieval using subspace clustering algorithm. In *Proceedings of ACM MMDB 04*, pages 100–108, 2004.

[16] J. Wu, M. Kankanhalli, J. Lim, and D. Hong. *Perspectives on Content-Based Multimedia Systems*. Kluwer Academic Publishers, 2000.

[17] R. Zhao and W. Grosky. Narrowing the semantic gap-improved text-based web document retrieval using visual features. *IEEE Transaction on Multimedia*, 4(2):189–200, 2002.