Detecting "In-Play" Photos in Sports News Photo Database

Akio Kitahara and Keiji Yanai

Department of Computer Science,
The University of Electro-Communications,
1-5-1 Chofugaoka, Chofu-shi, Tokyo, 182-8585, Japan
{yanai,kitaha-a}@mm.cs.uec.ac.jp

Abstract. In this paper, we treat with in-play classification of sports news photos as an instance of researches on more sophisticated search methods for large-scale photo news databases. We propose two methods to classify sports news photos into one of the given six sports categories and to discriminate in-play photos from not-in-play ones. One is the two-step method which classifies sports categories first and recognizes in-play conditions next, and the other is the one-step method which classifies them simultaneously. In the proposed methods, we integrate textual features extracted from news articles and image features extracted from photo images by Multiple Kernel Learning (MKL). In the experiment of the two-step method, we obtained 99.33% as the classification rate for the sports category classification which is the first step and 80.75% for the in-play classification which is the second step. On the other hand, in the experiment of the one-step method, we obtained 77.08% which was a little less than the result by the two-step method.

1 Introduction

Many commercial news sites exist on the Web, and they deliver a lot of new articles to us every day. Most of Web news articles contain photos, which help us understand the contents of the articles intuitively. Therefore, we can regard Web news database as integrated database of article texts and images. Since data on the Web can be collected automatically by crawling softwares, we can gather news articles on the Web every night automatically and build a personal news database with almost no cost. However, we can gather several thousands of articles a month, so that it is very difficult to watch all of them and find out interesting articles out of such huge news database.

In general, Web news articles are categorized only roughly into broad categories such as "sports", "politics", "economics" and "people life" on the Web news sites, and they are not classified in more detail categories such as "baseball" and "soccer" in advance. Moreover, photos included in sports news articles contain many types of scenes which represents playing a game, training, interview, press conferences and so on. Then, in this paper, we focus on discriminating in-play photos from not-in-play ones in addition to categorizing sports news photos into the given sports categories. Note that in this paper we define the

P. Muneesawang et al. (Eds.): PCM 2009, LNCS 5879, pp. 268–279, 2009.

photos which represents that players are playing in the sports field as being "in-play", and photos except "in-play" ones such as interview and press conference as being "not-in-play". If we can analyze the content of photos and articles automatically, more complicated search can become possible. For example, we will be able to search Web photo news database for an article including the photo which represents "Alex Rodriguez is hitting a ball in the game".

To classify news photos based on the content of photos, it is important to integrate both textual information and visual information. In this paper, as one case study of Web news classification with both image and text analysis, we classify sports news photos into one of the pre-defined sports categories and discriminate in-play photos from not-in-play ones. We propose two methods to do that. One is the two-step method which carries out sports category classification first and recognizes in-play conditions next, and the other is the two-step method which classifies them simultaneously. In the proposed methods, we integrate textual features extracted from news articles and image features extracted from photo images by the Multiple Kernel Learning (MKL).

We examined the effectiveness of our proposed method with Web photo news articles gathered from the Yahoo! Japan Photo News[1]. As sports categories, we prepared the following six categories: baseball, golf, F1, soccer, tennis, and sumo[2]. In the experiment of the two-step method, we obtained 99.33% as the classification rate for the sports category classification which is the first step and 80.75% for the in-play classification which is the second step. On the other hand, in the experiment of the one-step method, we obtained 77.08% which was a little less than the result by the two-step method.

The rest of this paper is organized as follows: In Section 2 we describe related work. In Section 3 we overview our approach, and in Section 4 we explain how to extract textual and image features from news article texts and photos and how to classify them. In Section 5 we presents the experimental results and evaluations, and in Section 6 we conclude this paper.

2 Related Work

Quattoni et al.[1] proposed a method to classify Web photo news articles into pre-defined genres, which integrates bag-of-features extracted from photos and bag-of-words extracted from article texts. Although they did not limit their method to sports news, they did not classify photos into in-play or not-in-play which are more precise categories than news genres.

As work related to classification of sports photos, Jain et al. [2] collected consumer sports photos most of which are taken by people from the seat of stadiums, and classify them into one of the pre-defined sports categories by recognizing sports fields. Their target is consumer sports photos, while ours is sports new photos taken by professional photographers.

As another work of sports photo classification, Li et al.[3] proposed a method to classify photos gathered from not Web news but general Web into sports categories by integrating scene recognition methods of whole images and object

[1] http://headlines.yahoo.co.jp/hl?ty=p

[2] Japanese traditional-style wrestling

recognition methods. The difference to this paper is that they classify sports photos into sports categories and does not classify photos as in-play or not-in-play which are more precise categories than sports categories.

3 Overview

In this paper, we propose two methods to classify sports news photos into one of the pre-defined sports categories and to classify them as in-play or not-in-play. One is the two-step method which classifies them into sports categories first and recognizes in-play conditions next, and the other is the one-step method which classifies sports news photos in terms of both sports categories and in-play condition simultaneously.

In the two-step method, the processing consists of the first step for sports category classification by text analysis and the second step for in-play classification by image analysis. In the first step, we extract textual features from news article texts to classify Web sports articles into sports categories. We use the bag-of-words representation which is based on the vector space model as features, and classify them using a Support Vector Machine (SVM) with the one-vs-rest strategy as a multi-class classifier. In the second step, we classify each sports photo as in-play or not-in-play with image features. As image features, we use the bag-of-features (BoF) representation for a whole image, the bag-of-features representation for regions (region-based BoF), the Gabor texture feature and a color histogram, and integrate them by using the Multiple Kernel Learning (MKL). MKL is a kind of extensions of SVM, and it enables to estimate optimal weights of different features and integrate them with the weighted sum of kernels. The detail of MKL will be described in the next section. Integrating whole-image bag-of-features and region-based bag-of-features by the Multiple Kernel Learning has not done so far. This is the first attempt of MKL-based integration of both of them.

In the one-step method, we classify photos into sports categories and in-play conditions simultaneously. The method is similar to the second step of the two-step method. In this method, we integrate textual features as well as four kinds of image features for classification by the Multiple Kernel Learning. In the experiment, since we handle six sports categories and two in-play conditions (in-play or not-in-play), totally twelve classes are prepared and we carry out twelve-class classification using the one-vs-rest strategy.

4 The Proposed Method

In this paper, we assume that articles of Web photo news consist of both texts and photos. Therefore, we can extract both textual and visual features from articles. We use the bag-of-words representation as textual features, and bag-of-features (BoF) representation for a whole image, the bag-of-features representation for regions (region-based BoF), the Gabor texture feature and a color histogram as image features.

To integrate them, we use the Multiple Kernel Learning (MKL) which can estimate optimal weights for linear combination of various features.

4.1 Bag-of-Words as Textual Features

Bag-of-words (BoW) is the standard representation for textual data. In the bag-of-words representation, documents are regarded as sets of words. To extract bag-of-words vectors, we count the frequency of words ignoring the order of words. Although it is easy to extract bag-of-words features from documents, it is known to be very effective in terms of document classification. In the proposed methods, bag-of-words features are used in the first step of the two-step method, and are used in the one-step method as one of the features to be integrated.

The procedure to extract bag-of-words vectors from article texts is as follows:

1. Count the frequency of words after removing stop words
2. Build a codebook by selecting the top N frequent words over all the news articles.
3. Generate a bag-of-words vector for each article by counting frequent words which appear in the codebook

4.2 Normal Bag-of-Features

The main idea of the bag-of-features (BoF) representation [4] is representing images as collections of independent local patches, and vector-quantizing them as histogram vectors.

The main steps of the method are as follows:

1. Sample 2000 patches per image randomly.
2. Generate feature vectors for the sampled patches by the SIFT descriptor [5].
3. Construct a codebook with k-means clustering over all the extracted feature vectors. A codebook is constructed for each concept independently. We set k as 1000 in the experiment.
4. Assign all SIFT vectors to the nearest codeword of the codebook, and convert a set of SIFT vectors for each region into one k-bin histogram vector regarding assigned codewords.

4.3 Region-Based Bag-of-Features

In addition to the normal BoF, we use region-based bag-of-features (region-based BoF). The difference to normal whole image BoF is that we apply region segmentation for images, before constructing bag-of-features vectors. As a region segmentation method, we use JSEG [6] after adjusting the parameters so as to generate about eight regions per image on average.

After applying region segmentation, we obtain several regions. We select the most relevant region to represent the given sports with a multiple instance learning method, and we use only one region-based BoF vector.

To select the most relevant region to the concepts, we use a multiple instance SVM (mi-SVM) classifier [7]. The mi-SVM is a support vector machine modified for multiple instance setting, and it is carried out by iterating a training step and a classification step using a standard SVM.

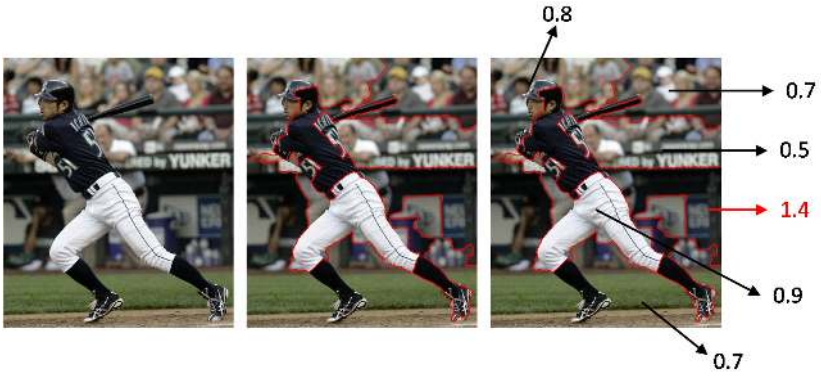

Fig. 1. Region selection by mi-SVM. Left: original image, center: segmented image, right: selected region by mi-SVM.

Under the multiple instance setting, training class labels are associated with a set of instances instead of individual instances. A positive set, which is called as a "positive bag", has one positive instance at least, while a negative set, which is called as a "negative bag", has only negative instances. This multiple instance setting fits well with the situation where an image consists of several foreground regions and background regions. Since we can regard foreground and background regions as positive and negative instances, respectively, by using multiple-instance learning methods we can classify regions into foregrounds and backgrounds.

At the first iteration, we prepare training images and random negative images as positive bags and negative bags, respectively. Next, we train a standard SVM treating with all the regions of positive bags as positive training samples and all the regions of negative bags as negative training samples. After that, we apply the trained SVM to all the vectors of all the regions of positive bags and classify the regions of positive bags into positive regions or negative regions.

At the second iteration, we use only regions classified into positive ones in the first step as positive samples and all the other regions as negative samples, and train the SVM again. Finally we apply the trained SVM to all the regions of candidate images, and obtain the output value of the SVM for each region which corresponds to the distance between the given vector and the discriminative hyper-plane in the context of SVM. We select the most relevant region according to the output values. Figure 1 shows an example of the output values of the mi-SVM.

4.4 Gabor Features

The Gabor texture feature represents texture patterns of local regions with several scales and orientations. In this paper, we use 24 Gabor filters with four kinds of scales and six kinds of orientations. Before applying the Gabor filters to an image, we divide an image into 4×4 blocks. We apply the 24 Gabor filters to each block, then average filter responses within the block, and obtain a

24-dim Gabor feature vector for each block. Finally we simply concatenate all the extracted 24-dim vectors into one 384-dim vector for each image.

4.5 Color Histogram

A color histogram is a very common image representation. We divide an image into 4×4 blocks, and extract a 64-bin RGB color histogram from each block with dividing the space into $4 \times 4 \times 4$ bins. Totally, we extract a 1024-dim color feature vector from each image.

4.6 Support Vector Machine

We use a standard Support Vector Machine (SVM) as a classifier for the first step of the two-step method which requires only bag-of-words vectors as features. SVM is basically a two-class classifier. For multi-class classification, we need to combine several SVMs, and we adopt the one-vs-rest strategy. In the experiment, we build six kinds of sports category detectors by regarding one category as a positive set and the other five categories as negative sets.

As a kernel for SVMs, we use the following chi-square RBF kernel:

$$K(a,b) = \exp(-\gamma * \sum_{i=1}^{n} \frac{|a_i - b_i|^2}{a_i + b_i}) \tag{1}$$

4.7 Multiple Kernel Learning

To integrate various kinds of image features and textual features in the one-step method and in the second step of the two-step method, we use a Multiple Kernel Learning (MKL). MKL is a kind of extensions of a support vector machine (SVM). MKL treats with a combined kernel which is a weighted liner combination of several single kernels, while a standard SVM treats with only a single kernel. MKL can estimates optimal weights for a linear combination of kernels as well as SVM parameters simultaneously in the train step. The training method of a SVM employing MKL is sometimes called as MKL-SVM. Since MKL-SVM is a relatively new method which was proposed in 2003 in the literature of machine learning [7], there have been not many works which applied MKL into image recognition so far.

With MKL, we can train a SVM with a adaptively-weighted combined kernel which fuses different kinds of image features. The combined kernel is as follows:

$$K_{comb}(\mathbf{x},\mathbf{y}) = \sum_{j=1}^{K} \beta_j K_j(\mathbf{x},\mathbf{y})$$
$$\text{with } \beta_j \geq 0, \ \sum_{j=1}^{K} \beta_j = 1. \tag{2}$$

where β_j is weights to combine sub-kernels $K_j(\mathbf{x},\mathbf{y})$. As a kernel function, we used a chi-square RBF kernel (Equation 1).

By preparing one sub-kernel for each image features and estimating weights by the MKL method, we can obtain an optimal combined kernel. We can train a SVM with the estimated optimal combined kernel from different kinds of image features efficiently.

Sonnenburg et al.[8] proposed an efficient algorithm of MKL to estimate optimal weights and SVM parameters simultaneously by iterating training steps of a standard SVM. This implementation is available as the SHOGUN machine learning toolbox at the Web site of the first author of [8]. In the experiment, we use the MKL library included in the SHOGUN toolbox as the implementation of MKL.

5 Experiments

5.1 Experimental Settings

We have collected Yahoo! Japan Photo News for several years. We selected sports news articles related to six sports categories, which are baseball, golf, F1, soccer, tennis and sumo, from our news collection. We picked up 100 in-play photo articles and 100 non-in-play ones for each sports category from the collection gathered in 2007 as training data, and in the same way we picked up 100 in-play photo articles and 100 non-in-play ones for each category from the 2008 news collection as test data. The decision on being in-play or not-in-play for each photo was made subjectively. Basically, photos that players are playing in the field are classified into in-play photos. Totally, we selected 2400 photo news articles for experiments by hand. Note that raw articles in Yahoo! Japan Photo News are tagged with not sports categories such baseball and soccer but only news genres such as "sports", "economy" and "local".

In the experiments, we use the classification rate for evaluation which is represented by

(number of correctly classified articles) / (number of all the articles) .

5.2 Two-Step Method

Sports Category Classification. In the first step of the two-step method, sports category classification was carried out. We made experiments in case that the size of codebook is 1000 and 2800, respectively. 2800-word codebook consists of 1000 global frequent words and 300 category-specific frequent words for each of six sports categories, while 1000-word codebook consists of only 1000 global frequent words.

As results, we obtained 99.00% and 99.33% classification rate with the 1000- and 2800-word codebook, respectively. These results shows that textual bag-of-words representation has enough power for sports category classification. Table 1 shows confusion matrix in case of the 2800-word codebook. The results for golf, F1 and sumo was perfect.

Table 1. Confusion matrix with the 2800-word codebook

	baseball	golf	F1	soccer	tennis	sumo
baseball	195	0	4	0	0	0
golf	0	200	0	0	0	0
F1	0	0	200	0	0	0
soccer	0	0	0	199	0	1
tennis	1	1	0	0	198	0
sumo	0	0	0	0	0	200

In-play Classification. In the second step of the two-step method, we classified sports photos into in-play or not-in-play ones by the MKL. In the experiments, we estimated the best setting in terms of γ which is a kernel parameter of the chi-square RBF kernel (Equation 1) with cross-validation using only training data.

Table 2 shows results by using each single feature and integrating all of them by the MKL. In the table, "BoF", "region-BoF", "Gabor" and "color" represent the normal bag-of-features, the region-based bag-of-features, the Gabor texture features, and a color histogram, respectively. From these results, the best results were obtained in case of feature integration by the MKL for all the categories. In case of F1 and sumo, single features achieved relatively good results, while single features made poor results in case of golf and tennis. Regarding the average of the results, classification rate of the region-based BoF was the worst, while results of the Gabor features was the best. This is partly because relevant region selection by mi-SVM sometimes failed.

Figure 2 shows the weights estimated by the MKL to integrate all the visual features. Note that in case of F1, tennis and sumo, two combinations of the optimal weights were estimated. This is because the two values of γ, which achieved the best results in the cross-validation for the parameter estimation, were detected, respectively. The estimated weights are different for both value as shown in Figure 2.

Regarding F1, tennis and sumo, the weight of BoF and the weight of region-based BoF were almost the same, and either of them are selected exclusively. For these results, BoF and region-based BoF is essentially the same features.

Table 2. The classification rate of the in-play classification for each sports category and the average of six sports categories using single features and MKL

	BoF	region-BoF	Gabor	color	MKL
baseball	76.0%	71.5%	80.5%	72.0%	**81.5%**
golf	60.5%	58.0%	66.5%	67.5%	**67.5%**
F1	85.5%	82.5%	84.0%	86.0%	**91.0%**
soccer	74.5%	73.5%	73.0%	77.5%	**81.0%**
tennis	63.0%	57.5%	68.5%	65.0%	**71.0%**
sumo	87.0%	78.5%	91.0%	91.0%	**92.5%**
AVG.	74.42%	70.25%	77.25%	76.50%	**80.75%**

In case of golf, as shown in Table 2, MKL did not improve the result by a color histogram. Therefore, in case of using MKL for golf, the weight of the color histogram became 0.84, which is very large weight for a single feature. This means that color is important for in-play classification of golf photos.

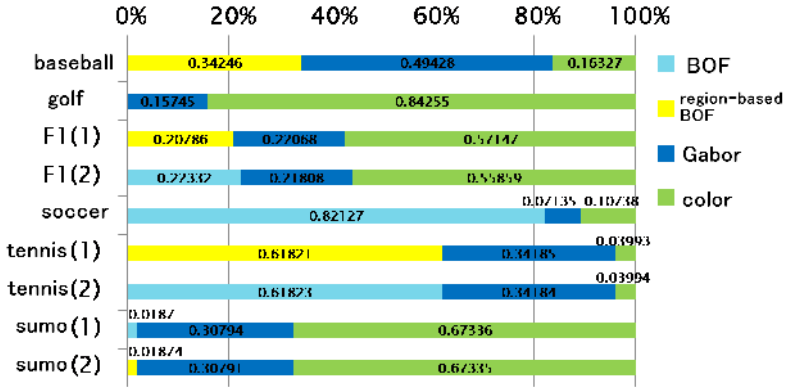

Fig. 2. Weights estimated by the MKL for the in-play classification

Fig. 3. Results of classification of F1 photos the first and second rows: true positive (in-play)the third rows: false positive, the fourth and fifth rows: true negative (not-in-play) the sixth row: false negative)

Fig. 4. Results of classification of Golf photos the first and second rowstrue positive (in-play)the third and forth rows: false positive, the fifth and sixth rows: true negative (not-in-play) the seventh and eighth rows: false negative)

The weight of BoF for soccer, 0.82, was also relatively large, which shows local pattern is important for in-play classification of soccer photos.

In case of sumo, the improvement by MKL was small, since even single Gabor features or color histograms achieved 91.0%.

We show some results of F1 photos in Figure 3, which achieved the 91.0% classification rate, and some results of golf photos in Figure 4, which achieved only 67.5%. Regarding golf photos, both in-play and not-in-play photos are taken in golf courses. This made it difficult to classify them.

5.3 One-Step Method

Table 3 shows the results of the one-step method which integrates all the visual features and textual features at the same time. The 77.08% classification rate was obtained on the average, which is a little less than the result of the two-step method, 80.75%. The single feature which brought the best result, 72.08%,

Table 3. The results by the one-step classification by single features and integrating all of the features by the MKL

feature	classification rate
BoW (text)	10.50%
BoF	72.08%
region-BoF	67.67%
Gabor	39.58%
color	49.17%
MKL	77.08%

Table 4. Confusion matrix by the one-step method **b**:baseball, **g**:golf, **f**:F1, **so**:soccer, **t**:tennis**su**:Sumo, **p**:in-play, **n**:not-in-play

	b_p	b_n	g_p	g_n	f_p	f_n	so_p	so_n	t_p	t_n	su_p	su_n
b_p	68	32	0	0	0	0	0	0	0	0	0	0
b_n	7	93	0	0	0	0	0	0	0	0	0	0
g_p	0	0	56	44	0	0	0	0	0	0	0	0
g_n	0	0	22	78	0	0	0	0	0	0	0	0
f_p	0	0	0	0	81	19	0	0	0	0	0	0
f_n	0	0	0	0	7	93	0	0	0	0	0	0
so_p	0	0	0	0	0	0	85	15	0	0	0	0
so_n	0	0	0	0	0	0	22	78	0	0	0	0
t_p	0	0	0	0	0	0	0	0	17	83	0	0
t_n	0	0	0	0	0	0	0	0	7	93	0	0
su_p	0	0	0	0	0	0	0	0	0	0	87	13
su_n	0	0	0	0	0	0	0	0	0	0	4	96

was normal BoF, and the difference between it and the result by the MKL corresponds to the improvement due to feature fusion.

Table 4 shows the confusion matrix of the one-step classification, which corresponds to twelve-class classification in the experiments. Although there were some confusions within the same sports category, there were no confusions across sports categories. This means that the classification rate in terms of sports category classification by the one-step method was 100.0%, while one by the two-step method was 99.33%. By integrating visual features as well as textual features, the results of sports category classification was improved. In addition, many confusions are found in the results of golf. Therefore, golf is a difficult category to classify in terms of in-play in case of the one-step method as well as the two-step method.

6 Conclusions

In this paper, we treated with in-play classification of sports news photos as an instance of researches on more sophisticated search methods for large-scale photo news databases. We proposed two methods to classify sports news photos into one of the pre-defined six sports categories and to discriminate in-play photos from not-in-play ones. One is the two-step method which carries out sports category classification first and recognizes in-play conditions next, and the other is the two-step method which classifies them simultaneously. In the proposed methods, we integrate textual features extracted from news articles and image

features extracted from photo images by Multiple Kernel Learning (MKL). In the experiment of the two-step method, we obtained 99.33% as the classification rate for the sport category classification which is the first step and 80.75% for the in-play classification which is the second step. On the other hand, in the experiment of the one-step method, we obtained 77.08% which was a little less than the result by the two-step method.

To improve classification performance, we need more various features which have different discriminative power from those of the currently-used features. For future work, we plan to introduce pose analysis of players in sports photos, and add new sports categories and new photo scene conditions such as interview and training in addition to in-play and not-in-play to the proposed system.

References

1. Quattoni, A., Collins, M., Darrell, T.: Learning visual representations using images with captions. In: Proc. of IEEE Computer Vision and Pattern Recognition (2007)
2. Jain, V., Singhal, A., Luo, J.: Selective hidden random fields: Exploiting domain-specific saliency for event classification. In: Proc. of IEEE Computer Vision and Pattern Recognition (2008)
3. Li, J., Li, F.: What, where and who? classifying events by scene and object recognition. In: Proc. of IEEE International Conference on Computer Vision (2007)
4. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Proc. of ECCV Workshop on Statistical Learning in Computer Vision, pp. 59–74 (2004)
5. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
6. Deng, Y., Manjunath, B.S.: Unsupervised Segmentation of Color-Texture Regions in Images and Video. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(8), 800–810 (2001)
7. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems, pp. 577–584 (2003)
8. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large Scale Multiple Kernel Learning. The Journal of Machine Learning Research 7, 1531–1565 (2006)