Recognizing Cultural Events in Images: a Study of Image Categorization Models

Heeyoung Kwon, Kiwon Yun, Minh Hoai, Dimitris Samaras
Stony Brook University, Stony Brook, NY 11794-4400
{heekwon, kyun, minhhoai, samaras}@cs.stronybrook.edu

Abstract

The goal of this work is to study recognition of cultural events represented in still images. We pose cultural event recognition as an image categorization problem, and we study the performance of several state-of-the-art image categorization approaches, including Spatial Pyramid Matching and Regularized Max Pooling. We consider SIFT and color features as well as the recently proposed CNN features. Experiments on the ChaLearn dataset of 50 cultural events, we find that Regularized Max Pooling with CNN, SIFT, and Color features achieves the best performance.

1. Introduction

Cultural events such as Rio Carnival and Munich Oktoberfest attract millions of visitors every year, and they are the attention of photographers, professionals and amateurs alike. As can be seen in Figure 1, the subjects of many consumer photographs are cultural events, and the ability to recognize those events in still images is useful for many applications such as indexing and searching of large-scale image archives.

The goal of this work is to study recognition of cultural events represented in typical still images such as consumer photographs. This problem has received little attention in the past, and it is only popularized by the recent ChaLearn Cultural Event Recognition Challenge [1].

We pose cultural event recognition as an image categorization problem and study the performance of several state-of-the-art image categorization approaches. In particular, we study the performance of Bag-of-Words representation combined with Spatial Pyramid Matching [12]. This is a popular approach which has been used for many tasks, including scene recognition [12] and human action recognition [5] in images. In addition to Spatial Pyramid Matching, we also investigate the recently proposed method Regularized Max Pooling [9]. This method has shown impressive state-of-the-art performance for recognizing human actions from still images.

The rest of this paper is organized as follows. Section 2 describes the ChaLearn Cultural Event dataset. Section 3 explains the feature representation used in our experiments. Sections 4 and 5 present two recognition methods investigated in this paper. Section 6 provides extensive experimental evaluation of different methods and parameter settings on the ChaLearn Cultural Event dataset.

2. Dataset and performance measure

Dataset. We study recognition of cultural events using the ChaLearn Cultural Event Recognition dataset [1]. This dataset consists of images of 50 cultural events over 28 countries. Examples of cultural events are Carnival Rio, Oktoberfest, St.Patrick's Day, La Tomatina, and Tango Festival. The images are collected by querying the image search engines of Google and Bing. The images are divided into three disjoint subsets for training, validation, and testing; these subsets contain 5,875, 2,332, and 3,569 images, respectively. Each image is manually annotated with one of the 50 cultural events. Figure 1 shows some example images for several cultural events such as Carnival Rio, Oktoberfest, and St Patrick's Day.

Performance measure. The performance measure is Mean Average Precision (mAP), which is a standard measurement used for benchmarking recognition systems in many visual recognition challenges including ChaLearn [1] and PASCAL VOC Challenges [7]. To obtain mAP, we first compute the Average Precision (AP) for each of the 50 cultural event classes, and the mAP is the average of the obtained APs. The AP for a class is calculated as the area under the precision-recall curve for the 1-vs-all cultural event classifier. The evaluation code is provided by ChaLearn [1]. Due to the unavailability of annotation for the test set, most results reported in this paper are performed on the validation data.

3. Classifiers and Features

To recognize cultural events, we use Least-Squares Support Vector Machines [19] and consider three types of local features. The local features are SIFT, Color, and CNN,

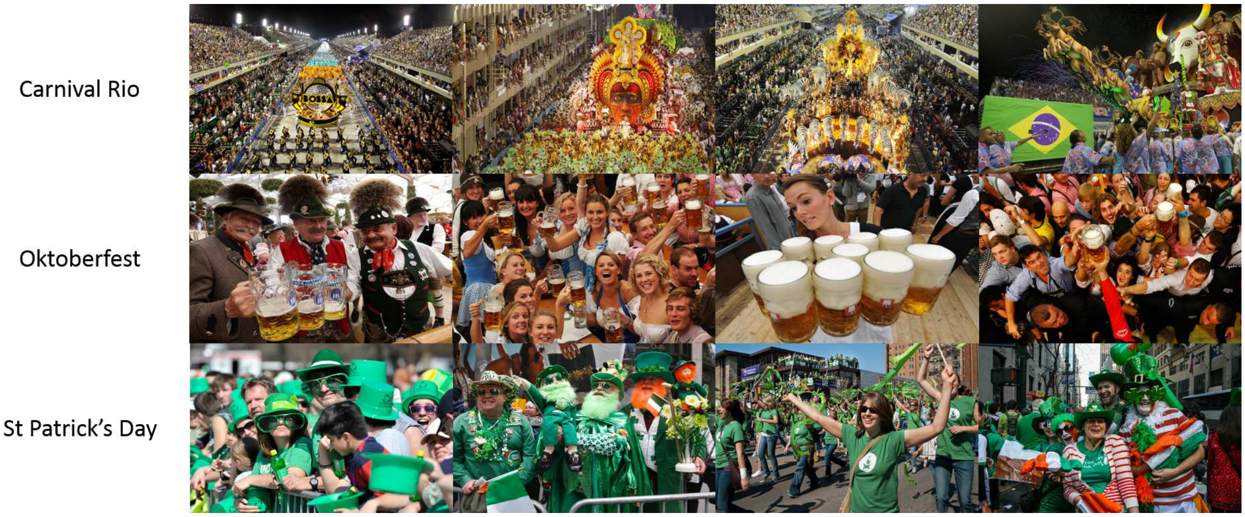


Figure 1. Example images of the cultural events from ChaLearn Cultural Event dataset [1].

which are among state-of-the-art features for image categorization. We additionally consider some combinations of these features.

3.1. Least-Squares SVM classifiers

We use Least-Squares Support Vector Machines (LSSVM) [19]. LSSVM, also known as kernel Ridge regression [16], has been shown to perform equally well as SVM in many classification benchmarks [18]. LSSVM has a closed-form solution, which is a computational advantage over SVM. Furthermore, once the solution of LSSVM has been computed, the solution for a reduced training set obtaining by removing any training data point can found efficiently. This enables reusing training data for further calibration (e.g., used in [9, 10]). This section reviews LSSVM and the leave-one-sample-out formula.

Given a set of n data points $\{\mathbf{x}_i | \mathbf{x}_i \in \Re^d\}_{i=1}^n$ and associated labels $\{y_i | y_i \in \{1, -1\}\}_{i=1}^n$, LSSVM optimizes the following:

$$\underset{\mathbf{w}, b}{\text{minimize}} \ \lambda ||\mathbf{w}||^2 + \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i + b - y_i)^2. \qquad (1)$$

For high dimensional data ($d \gg n$), it is more efficient to obtain the solution for $(\mathbf{w}, b)$ via the representer theorem, which states that $\mathbf{w}$ can be expressed as a linear combination of training data, i.e., $\mathbf{w} = \sum_{i=1}^n \alpha_i \mathbf{x}_i$. Let $\mathbf{K}$ be the kernel matrix, $k_{ij} = \mathbf{x}_i^T \mathbf{x}_j$. The optimal coefficients $\{\alpha_i\}$ and the bias term b can be found using closed-form formula: $[\boldsymbol{\alpha}^T, b]^T = \mathbf{M}\mathbf{y}$. Where $\mathbf{M}$ and other auxiliary variables are defined as:

$$\mathbf{R} = \begin{bmatrix} \lambda \mathbf{K} & \mathbf{0}_n \\ \mathbf{0}_n^T & 0 \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{K} \\ \mathbf{1}_n^T \end{bmatrix}, \qquad (2)$$

$$\mathbf{C} = \mathbf{R} + \mathbf{Z}\mathbf{Z}^T, \mathbf{M} = \mathbf{C}^{-1}\mathbf{Z}, \mathbf{H} = \mathbf{Z}^T\mathbf{M}. \qquad (3)$$

If $\mathbf{x}_i$ is removed from the training data, the optimal coefficients can be computed:

$$\begin{bmatrix} \boldsymbol{\alpha}_{(i)} \\ b_{(i)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} + \left(\frac{[\boldsymbol{\alpha}^T \ b]\mathbf{z}_i - y_i}{1 - h_{ii}} \right) \mathbf{m}_i. \qquad (4)$$

Here, $\mathbf{z}_i$ is the i^{th} column vector of $\mathbf{Z}$ and h_{ii} is the i^{th} element in the diagonal of $\mathbf{H}$. Note that $\mathbf{R}, \mathbf{Z}, \mathbf{C}, \mathbf{M}$, and $\mathbf{H}$ are independent of the label vector $\mathbf{y}$. Thus, training LSSVMs for multiple classes is efficient as these matrices need to be computed once. A more gentle derivation of the above formula is given in [3].

3.2. SIFT

We consider cultural event recognition using PHOW descriptors [2]. PHOW descriptors are dense SIFT [15] at multiple scales. Specifically, we extract dense SIFT at every 5 pixels at scales 4, 6, 8, and 10. To extract PHOW descriptors, we use VLFEAT library [20]. We build a visual vocabulary from training SIFT descriptors using k-means with $4,000$ clusters. Subsequently, SIFT descriptors are assigned to one of the clusters (i.e., visual words), and the descriptors for an image (or an image region in case of Spatial Pyramid Matching) are aggregated into a 4000-dimensional histogram, which is referred to as the bag-of-words representation.

3.3. Color

Although SIFT descriptors can be used to capture the texture of cultural events, they ignore color information. Color cues, however, are crucial for identifying some cultural events, as illustrated in Figure 1 (e.g., St Patrick's Day based on the green color). To capture color information, we divide an image into a grid of 3×3 cells, and compute the

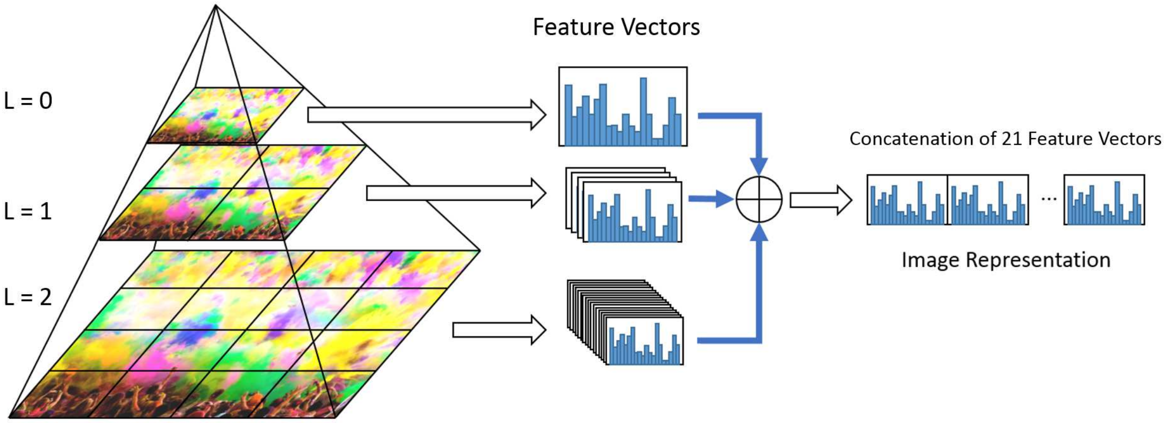


Figure 2. The pipeline for Spatial Pyramid Matching. An image is divided into 1×1 (Level 0), 2×2 (Level 1) and 4×4 (Level 2) spatial grids of cells. A local feature vector for each cell is computed, and all feature vectors are concatenated to create a feature vector for the image. Local feature vectors could be CNN features or the BoW representation of SIFT/color descriptors.

average normalized RGB values in each grid cell. Given the unnormalized (R, G, B) values, the normalized RGB values are defined as R/(R+G+B), G/(R+G+B), B/(R+G+B). The normalized RGB values are known as rg chromaticity [14], which are less sensitive to illumination conditions than the unnormalized RGB values. We build a color codebook using k-means with $2,000$ clusters. Subsequently, color descriptors are assigned to one of the color words, and the descriptors for and image (or an image region) are aggregated into a 2000-dimensional histogram.

3.4. CNN features

We also study the use of CNN features [13] for cultural event recognition. CNN features have been shown to yield excellent image classification results [11, 4]. To extract CNN features, we use the publicly available Caffe implementation [4] of the CNN architecture described by Krizhevsky et al. [11]. As will be seen, we extract CNN features for multiple regions of an image. For each region, we extract a 1024-dimensional CNN feature vector. This feature vector is referred to as CNN_M_1024 in [4]. As shown by [4], this feature is slightly worse than CNN_M_2048 and CNN_S_TUNE, but it has lower dimensionality.

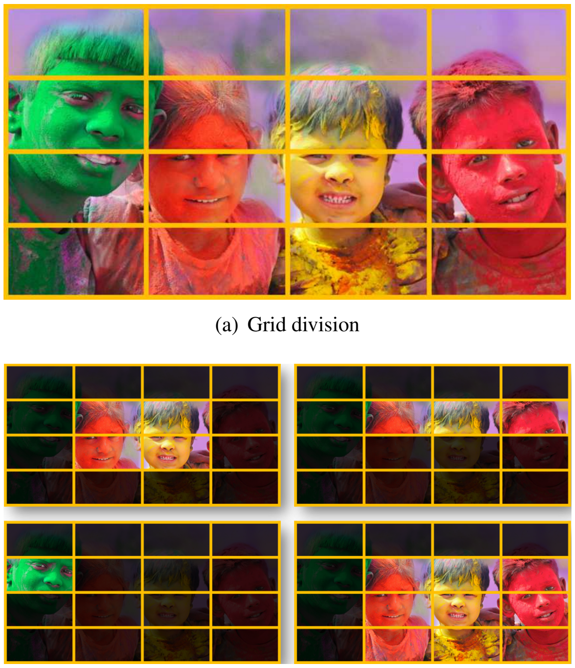
4. Spatial Pyramid Matching

Spatial Pyramid Matching (SPM) [12] is a popular approach for image categorization. SPM works by partitioning an image into increasingly fine sub-regions and aggregating local features found inside each sub-region (e.g., computing histograms [17] of quantized SIFT descriptors [15]). This approach has shown impressive levels of performance on various image categorization tasks. In our experiments, we consider a SPM model with three levels [12]. Specifically, an image is divided into 1×1 (Level 0), 2×2 (Level 1) and 4×4 (Level 2) spatial grids of cells. A feature vector is computed for each cell, and all feature vectors are concatenated with weights 0.25, 0.25, 0.5 for levels 0, 1, 2, respectively. This yields a $(1 + 4 + 16)K = 21K$ dimensional feature vector, where K is the dimension of local feature vectors for cells. Figure 2 illustrates the pipeline of SPM.

We use SPM with all three types of local features: SIFT, color, and CNN. For SIFT and Color, the local feature vector for each cell is the Bag-of-Words (BoW) representation [17] of visual words. The vocabulary sizes for SIFT and Color codebooks are $4,000$ and $2,000$ respectively.

5. Regularized Max Pooling

Regularized Max Pooling (RMP) [9] is a recently proposed technique for image categorization. It has been shown to achieve state-of-the-art performance on recognizing human actions in still images. RMP combines the flexibility of Spatial Pyramid Matching (SPM) and the deformability of Deformable Part Model (DPM) [8]. RMP works by partitioning an image into multiple regions and aggregating local features computed for each region, as in the case of SPM. However, unlike SPM, RMP does not rely on rigid geometric correspondence of grid division. Grid division ignores the importance of semantic or discriminative localization; and it has limited discriminative power for recognizing semantic category with huge variance in location or large deformation. To overcome the rigidity of grid division, RMP allows the regions to deform, just like in the case of DPMs with deformable parts. The regions in an RMP can be considered as parts at different locations and scales. Parts are geometrically anchored, but can be dis-



(a) Grid division

(b) Examples of considered subwindows

Figure 3. **From grid division to subwindows**. (a): An image is divided into 4×4 blocks. RMP considers rectangular subwindows that can be formed by a contiguous chunk of blocks. There are 100 such subwindows, and (b) shows four examples.

criminatively deformed. Compared to DPM, RMP has two advantages. First, an RMP model can have hundred of parts at various scales while a DPM is typically used with a small number of parts at the same scale. Second, the learning formulation of RMP is simple, without the need for expensive iterative updates.

An RMP model is a collection filters. Each filter is anchored to a specific image subwindow and associated with a set of deformation coefficients. The anchoring subwindows are predetermined at various locations and scales, while the filters and deformation coefficients are learnable parameters of the model. Figure 3 shows how the subwindows are defined. To classify a test image, RMP extracts feature vectors for all anchoring subwindows. The classification score of an image is the weighted sum of all filter responses. Each filter yields a set of filter responses, one for each level of deformation. The deformation coefficients are the weights for these filter responses. Please refer to [9] for more details.

6. Results

This section reports the performance of various methods, considering different combinations of feature types and image categorization models. We first discuss the performance of SPM using SIFT, color, and CNN features. We will then compare the performance of SPM and RMP.

Row	Feature type	Method type	mAP
(A)	SIFT	SPM	47.8
(B)	SIFT+Color	SPM	55.8
(C)	CNN	SPM	70.1
(D)	CNN	MultiReg	70.3
(E)	CNN	RMP	71.9
(B) + (E)			**73.7**

Table 1. Performance of various methods with different combination of feature types and methods. Color feature provides complementary cues for SIFT, and CNN performs better than both SIFT and Color combined. RMP outperforms MultiReg and SPM. The best result is achieved by combining SPM using SIFT+Color and RMP using CNN.

6.1. Feature Comparison

The first three rows of Table 1 show the performance of SPM using SIFT, SIFT+Color, and CNN features. As can be seen, all feature types perform reasonably well, achieving high mean average precisions for 50 cultural event classes. SIFT+Color outperforms SIFT by a large margin; this confirms the benefits of using color cues for recognizing cultural events. Overall, CNN features perform the best, achieving mAP of 70.1%. For a reference, CNN with a single layer SPM (Level-0 only) achieves mAP of 64.5%.

6.2. Method comparison

Rows (C), (D), and (E) of Table 1 report the performance of several methods using the same feature type, CNN. We will refer to these methods as SPM, MultiReg, and RMP, hereafter. SPM is the method that uses 3-level spatial pyramid of CNN features. MultiReg is similar to SPM. It is also based on rigid geometric correspondence, but extracts features from a different set of subwindows. In particular, MultiReg divides an image into a grid of 16 blocks (4×4) and considers all 100 rectangular subwindows that can be obtained by a contiguous set of blocks (Figure 3). RMP uses the same set of subwindows as MultiReg, but the subwindows can deform (translation and scale). SPM, MultiReg, and RMP are all based on LSSVM. As can be seen, RMP outperforms SPM and MultiReg. Since all of these methods use the same feature type, the superiority of RMP can be accounted for by its ability to handle deformation.

We consider the differences between APs of RMP and SPM, which will be referred to as AP gaps. We sort the AP gaps for 50 cultural event classes and plot them in Figure 4. As can be seen, the AP gaps are positive for 75% of the classes. Figure 5 shows some representative images from two classes with the highest AP gaps and the bottom two classes with the lowest AP gaps. For classes in Figures 5(a) and 5(b), the AP gaps between RMP and SPM are 10.8%

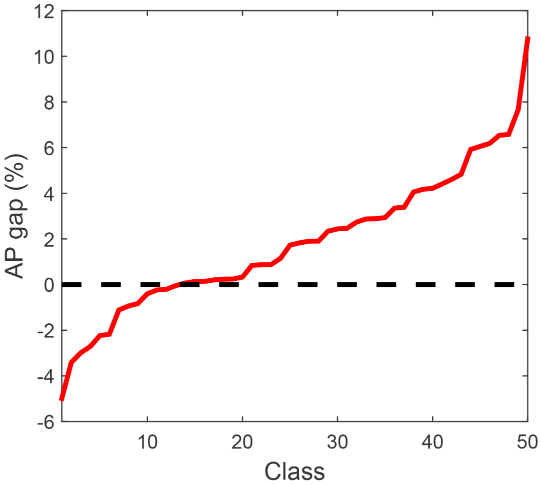


Figure 4. **AP gaps between RMP and SPM.** AP gaps are the differences between the APs of RMP and SPM, using CNN features. This figure plots the individual AP gaps for 50 cultural event classes. The majority of classes have a positive AP gap.

and 7.7% respectively. However, for classes in Figures 5(c) and 5(d), RMP does not perform better than SPM. The key advantage of RMP over SPM is the ability to handle geometric deformation. However, the intraclass variance of images in some cultural classes might not be due to geometric deformation, as can be seen in Figures 5(c) and 5(d).

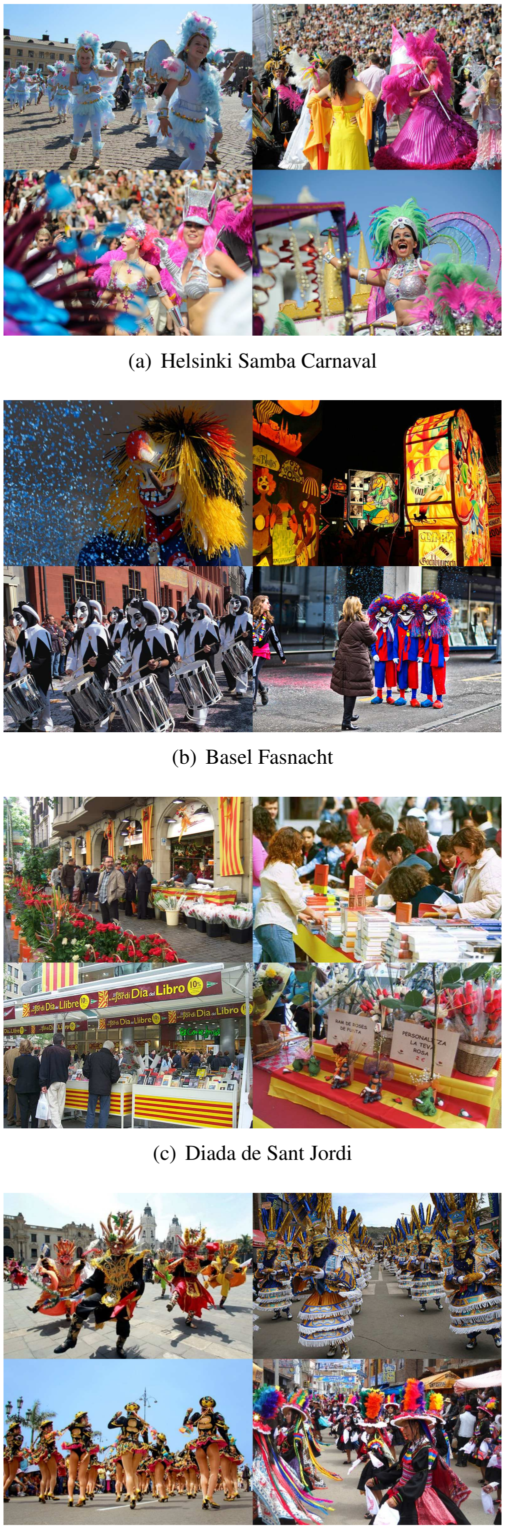
6.3. Method and feature combination

We also study the complementary benefits of SIFT+Color features to RMP, which uses CNN features. In particular, we combine SPM using SIFT+Color (Section 6.1) and RMP that uses CNN features as follows. To combine two methods, we consider the average of the SVM confidence scores. The mAP for the combined method is **73.7%**. This is significantly higher than 71.9%, which is the mAP of RMP alone. This suggests that SIFT+Color features provide complementary information for recognizing cultural events.

Figure 6 shows the APs of 50 event classes, obtained using a method that combines RMP using CNN and SPM using SIFT+Color. The majority of classes have APs higher than 60%. The two classes with highest APs are shown in Figure 7, and the two classes with lowest APs are shown in Figures 8 and 9. As can be seen, classes with low APs are not as distinctive as classes with high APs.

6.4. Comparison with other ChaLearn participants

We compare the performance of our best method (RMP+CNN combined with SPM+SIFT+Color) with the entries of ChaLearn 2015 Challenge. The results on the test set are obtained by submitting the output of our method to



(a) Helsinki Samba Carnaval

(b) Basel Fasnacht

(c) Diada de Sant Jordi

(d) Fiesta de la Candelaria

Figure 5. Cultural event classes with highest (a, b) and lowest (c, d) AP gaps between RMP and SPM.

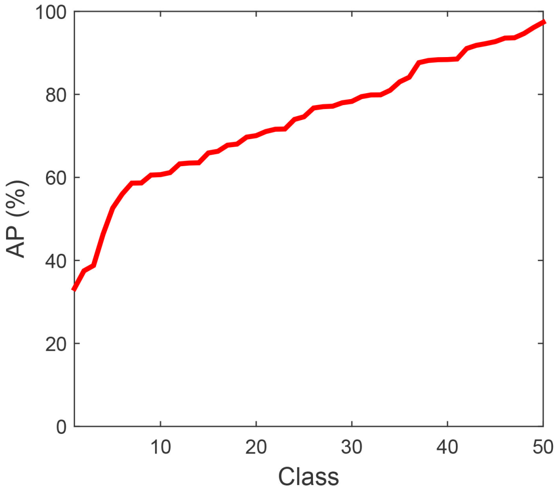
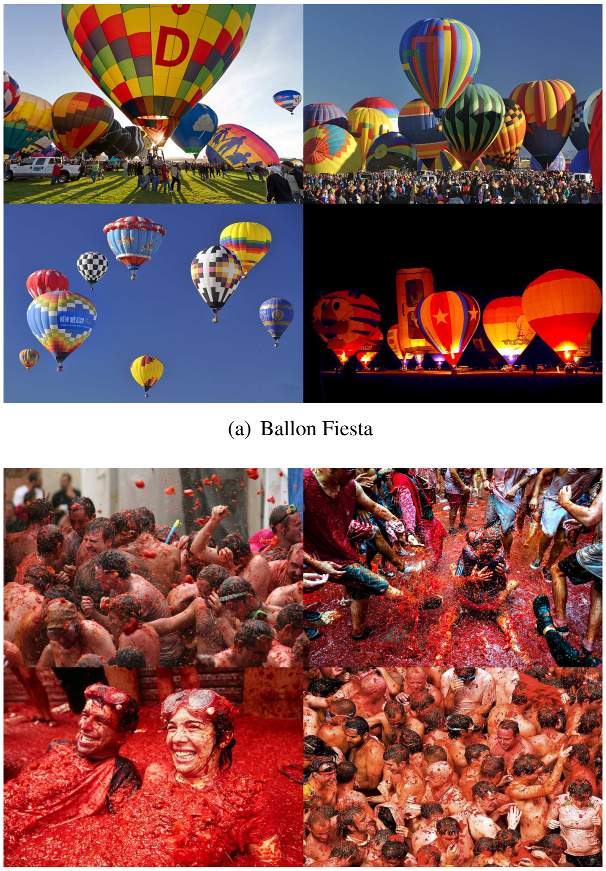
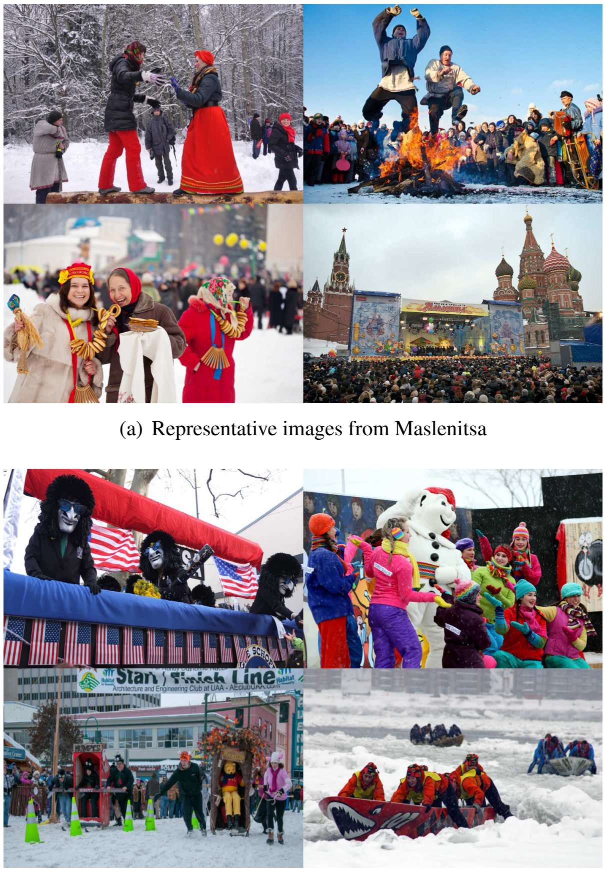


Figure 6. **Individual APs for 50 cultural classes.** These APs are obtained using a method that combines RMP using CNN and SPM using SIFT+Color. The mAP is 73.7. Only few classes have APs lower than 60%.


(a) Ballon Fiesta

(b) La Tomatina

Figure 7. Cultural event classes with highest APs (more than 96%). The images from these classes are highly distinctive.


(a) Representative images from Maslenitsa

(b) Negative images that receive high confidence scores

Figure 8. **Maslenitsa – cultural event with lowest AP**. (a): representative images of the class. (b): negative images with highest confidence. Images from this class are not as distinctive as images from classes shown in Figure 7.

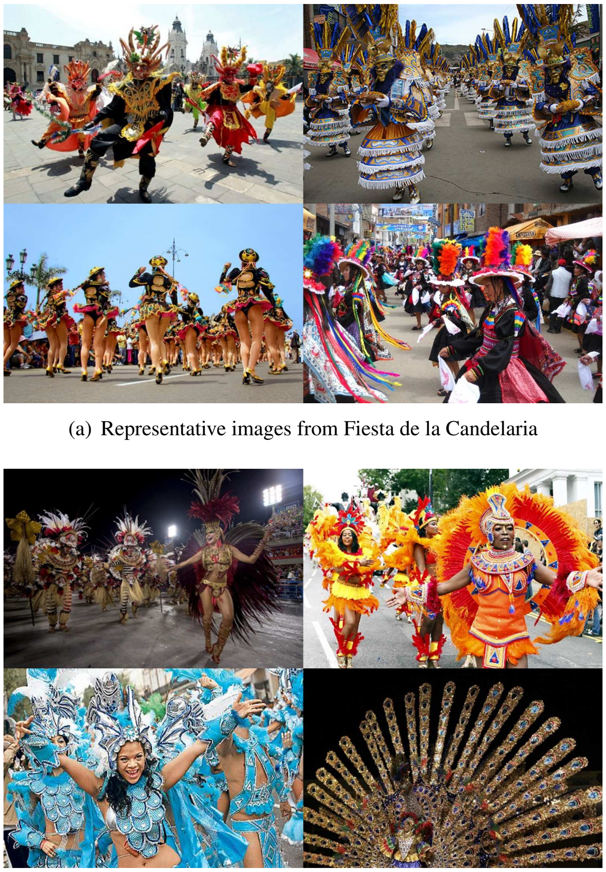
Team	mAP
Nyx	31.9
MasterBlaster	58.2
MIPAL_SNU	73.5
UPC-STP	76.7
MMLAB	85.5
Ours	74.2

Table 2. **Comparison with other entries of the 2015 ChaLearn Challenge on recognizing cultural events.**

the challenge organizer. This submission is done once. Table 2 shows the results of our method and other participants. Our method achieves the third best performance.

7. Conclusions

We have studied the performance of Spatial Pyramid Matching (SPM) and Regularized Max Pooling (RMP) on



(a) Representative images from Fiesta de la Candelaria

(b) Negative images that receive high confidence scores

Figure 9. **Fiesta de la Candelaria – cultural event with second lowest AP**. (a): representative images of the class. (b): negative images with highest confidence; these images are highly similar to actual images of the event class.

the task of cultural event recognition. We have considered several types of features and performed experiments on the ChaLearn dataset [1] that contains 50 cultural events. We observe that CNN features outperform SIFT and Color features, but SIFT and Color do provide complementary benefits. RMP outperforms SPM for the majority of cultural event classes. We visually examine the classes in which RMP does not lead to better performance and find that those classes contain images with high intraclass variance, but the variance is not mainly due to geometric deformation. In this work, we have used CNN_M_1024, which is an off-the-shelf CNN feature [4] trained on ImageNet dataset [6]. We have not finetuned CNN features on the ChaLearn Cultural Event dataset. We leave this for future work and expect it would improve the overall performance.

Acknowledgement: This work was supported in part by NSF grants IIS-1161876, IIS-1111047 and the SubSample Project by the DIGITEO institute France.

References

[1] X. Baro, J. Gonzalez, J. Fabian, M. A. Bautista, M. Oliu, I. Guyon, H. J. Escalante, and S. Escalers. Chalearn looking at people 2015 cvpr challenges and results: action spotting and cultural event recognition. In *CVPR*. ChaLearn Looking at People workshop, 2015.

[2] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *ICCV*, 2007.

[3] G. C. Cawley and N. L. Talbot. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Networks*, 17:1467–1475, 2004.

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[5] V. Delaitre, I. Laptev, and J. Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *BMVC*, 2010.

[6] J. Deng, W. Dong, R. Socher, K. L. L.-J. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE PAMI*, 32(9):1627–1645, 2010.

[9] M. Hoai. Regularized max pooling for image categorization. In *BMVC*, 2014.

[10] M. Hoai and A. Zisserman. Improving human action recognition using score distribution and ranking. In *ACCV*, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178. IEEE, 2006.

[13] Y. LeCun, B. Boser, J. S. Denker, and D. Henderson. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[14] M. D. Levine and M. D. Levine. *Vision in man and machine*, volume 574. McGraw-Hill New York, 1985.

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[16] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *ICML*, 1998.

[17] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[18] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. DeMoor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

[19] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[20] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.