A POOL OF DEEP MODELS FOR EVENT RECOGNITION

K. Ahmad, M. L. Mekhalfi, N. Conci, G. Boato, F. Melgani, F. G. B. De Natale

Department of Information Engineering and Computer Science, University of Trento, Italy

ABSTRACT

This paper proposes a novel two-stage framework for event recognition in still images. First, for a generic event image, deep features, obtained via different pre-trained models, are fed into an ensemble of classifiers, whose posterior classification probabilities are thereafter fused by means of an order-induced scheme, which penalizes the yielded scores according to their confidence in classifying the image at hand, and then averages them. Second, we combine the fusion results with a reverse matching paradigm in order to draw the final output of our proposed pipeline. We evaluate our approach on three challenging datasets and we show that better results can be attained, advancing recent leading works.

Index Terms— Event recognition, score-level fusion, deep features, score penalization.

1. INTRODUCTION

Over the last few years, event-based concepts have been heavily exploited for multimedia indexing and retrieval. Consequently, a number of interesting solutions have been proposed for event recognition in both videos and static images [1]. Given the dynamic nature of events, event recognition in images is deemed a more challenging task as we have less chromatic/spatial information in hand. Thus, conventional approaches relying on shallow handcrafted visual features cannot handle the semantic gap between the chromatic content of an image and its semantic attributes [2, 3].

Recently, Convolutional Neural Networks (CNN) have demonstrated promising performance in event recognition (e.g., [4, 5, 6, 7, 8]), thanks to the processing facilities which are more affordable than for instance just a decade ago. However, the existing datasets for event recognition are not large enough to satisfy the training requirements of CNNs. Thus, most of the existing approaches tend to fine-tune pre-trained CNNs to fit them into event related images [9, 5, 10].

Capitalizing on the success of CNNs trained on scene datasets on the one hand and object datasets on the other, recent works jointly couple the earlier two models for event recognition (e.g., [10, 11]). This blend of object-scene information has shown to improve, by far, the performance of event recognition frameworks. However, such simplistic fusion schemes treat both types of features equally, which is a naive approach provided that images manifest different spectral as well as spatial characteristics (i.e., some images may favor scene-level information over object features, and vice versa). To this end, we believe that, both the scene and object CNN models can be properly fused whilst preserving the merit of either model. This is achievable by assigning weights to the classification scores based on their confidence in predicting the class of a given event image. However, the existing weighting trends often consider learning a set of weights to be assigned to feature descriptors, which raises questions while switching from one dataset to another (i.e., the weights are dataset-specific, and thus need to be re-learned once the dataset is changed).

In this paper, we propose a novel fusion scheme, inspired by the concept of Induced Ordered Weighted Averaging Operators (IOWA) [12], to better utilize object-scene level information in event recognition. The underlying insight of the proposed IOWA fusion is to infer, in a learning-free fashion, weights from the decision scores yielded by a bunch of classifiers trained on different types of CNN features. The weights are thereafter utilized to penalize the individual decisions made by each classifier individually. Thus, the model that draws the most confident decision for a given event image is assigned the highest weight. Ultimately, the IOWA ends up by averaging the weighted classification scores in order to draw a single ultimate decision about the class of the image at hand, and we show that very prominent performances can be attained.

The second contribution of this paper is based on a reverse matching concept, which was proven to be useful in other computer vision tasks [13, 14]. Precisely, given a test image and an ensemble of training images, the proposed reverse matching-driven strategy proceeds by exchanging the test image with a bundle reference training images of each class iteratively, each time treating the training bundle as a test subject while the test image serves as a part of the training set. Thus, if the training and the test subjects belong to the same class, the test image must exhibit the lowest distance among all the training set, as elaborated further. The classification scores obtained at each iteration are finally aggregated and fused with the classification estimates yielded by the IOWA stage, which makes the final decision of the proposed event recognition pipeline. To the best of our knowledge, these two paradigms have never been proposed for event recognition.

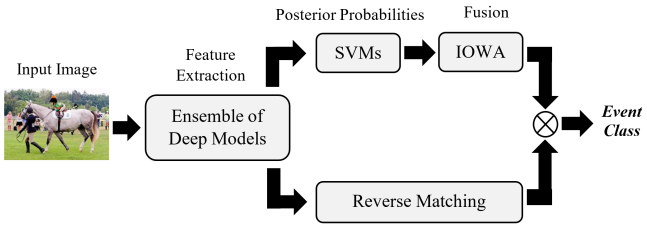

Fig. 1. Pipeline of the proposed framework.

We show that plausible results can be scored with respect to reference works. Besides the earlier two contributions, we also investigate the performance of three frequently used CNN architectures, with object and scene level information, in the event recognition context, which is also another key-contribution to be considered as a departing point for other researchers.

2. METHODOLOGY

As illustrated in Figure 1, the proposed framework for event recognition in static images proceeds by extracting a bunch of feature vectors via an ensemble of CNN models. Subsequently, the second phase is made up of two separate but integral parts. The first one takes in the extracted CNN features and feeds them into an ensemble of trained Support Vector Machines (SVMs), which in turn output posterior classification probabilities, which are finally fused via IOWA strategy in order to draw ultimate classification scores for a given event image. On the other side, the reverse ranking block takes in the CNN features, and performs a reverse matching as detailed later, which turns out to produce distance measures. The final phase fuses the outputs pointed out by the two earlier algorithms in order to make a final decision. Below are details outlining each phase.

2.1. Feature Extraction and Classification

Previous literature has evidenced the superiority of CNNs over traditional feature extraction schemes. However, two drawbacks can be observed when it comes to event recognition. On the one hand, the comparisons of various CNN architectures in event recognition is still missing. On the other hand, the mainstream approaches rely mostly on a single architecture pre-trained on either object or places or on both datasets. In order to mitigate the earlier two gaps, we investigate in this work the performance of three frequently used CNN architectures, namely Alexnet [15], GoogleNet [16] and VGG-16 [17] in event recognition. Alexnet [15] totals 8 weighted layers (5 convolutional and 3 fully connected layers), GoogleNet [16] contains 22 layers, whilst VGGNet-16 is composed of 16 layers. Moreover, in order to enrich the characterization of event images, we investigate the performance of different combinations of the earlier three architectures that are pre-trained on both objects [18] and places [19] datasets, which we believe is the first attempt in the event recognition context. Yet, six feature vectors (two per CNN architecture) are envisioned.

2.2. Order-Induced Posterior Probability Fusion

As elaborated in the introduction, unlike usual score-level fusion strategies, we put forth a learning-free fusion scheme that is inspired by a method known as Induced Ordered Weighted Averaging Operators (IOWA) by Yagr et al. [12, 20]. In designing our scheme, we depart from the fact that different CNN models entail different classification scores for a given set of event images. Hence, the wisest way to fuse their outcomes is to attribute high weights to the most certain CNNs, while assigning small weights to the less confident models, which forms the essence of our IOWA fusion method.

Suppose *N* is the number of different pre-trained networks, and that for each network an SVM classifier is assigned. If the total number of event classes in the database is represented by *M*, then, we get an *NxM* matrix, accommodating the posterior probabilities of an image with respect to all classes. Suppose p_i, $i = 1, 2, 3, ..., N$, is the score vector pertaining to the scores of the i^{th} classifier. The fusion strategy is aimed at gathering an ensemble of pairs $[p_i, o_i]$. Here p_i represents the argument value while o_i is its associated order-inducing value, which quantifies how confident the score p_i made by i^{th} classifier is. The IOWA operator, which represents the final outcome of the algorithm as the weighted sum of the reordered probabilities vectors, is given by:

$$F(p_i, o_i) = \frac{1}{N} \sum_{i=1}^{N} w_i s_i \tag{1}$$

where $W = [w_i, w_2, ..., w_N]$ represents the corresponding weights and $S = [s_1, s_2, ..., s_N]$ corresponds to $P = [p_1, p_2, ..., p_N]$ in a descending order (i.e., posterior probabilities arrays are sorted according to their associated inducing values o_i). In Algorithm 1 we provide a procedure for the selection of these weights.

As per the selection of the inducing value (i.e., o_i), which represents the confidence of its associated argument p_i, we opt for the standard deviation of the highest values in the array of classification probabilities p_i made by i^{th} classifier in order to treat the uncertainties in the highest scores. Thus, the event class with closest similarity corresponds to the more probable outcome in F. A step-wise elaboration of the proposed fusion scheme is shown in Algorithm 1, where w_i represents the new weights, and T_i is used to supress these weights gradually.

2.3. Reverse Matching Strategy

In order to compute reverse matching scores, in this work we randomly select 30 training images from each category as

Algorithm 1 Calculate IOWA fused score

Require: p_i and an inducing mechanism defining o_i.
Ensure: Fused decision vector F.
for i=1:N **do**
Step 1: Sort the pairs $< p_i, o_i >$ in a descending order based on their confidence measure o_i. Let $< s_i, v_i >$ be the reordered pairs.
Step 2: Set $r_0 = 1$ and $r_i = v_i$ for $(i = 1, ..N)$.
Step 3: Compute $T_i = \prod_{j=1}^{i} r_{j-1}$ (for $i = 1, ..N$).
Step 4: Infer the Weights $w_i = \frac{T_i}{\sum_{j=1}^{N} T_j}, (for$i=1,..N).
Step 5: Discern the ultimate classification score F according to equation 1.
end for

Algorithm 2 Calculate reverse matching scores

Require: A test image t_i and training set.
Ensure: Reverse score vector S.
for j=1:M **do**
Step 1: Swap t_i with the j^{th} category in training set.
Step 2: Compute Hausdorff distance between new test object j^{th} category and each training sample.
Step 3: Sort the distance vector. Let d_i be the reordered distance vector.
Step 4: Find the position of the t_i in the ranked list d_i.
Step 5: Compute the score of t_i, w.r.t j^{th} category (acting as a new test object), based on its position in the ranked list.
end for

representative images/collection. Subsequently, the test image is exchanged with the representative images of each category, iteratively. Thus, each time the representative images are treated as a test subject while the original test image is deemed as a part of the training set composed of the representative images/collections from other categories. For the similarity measurement between the new test subject (representative set of a category) and the training samples, including the original test image, we rely on a bag-level distance, the modified Hausdorff Distance, which has been successfully adopted in a number of works for the similarity measurement among bags in Multiple instance learning scenarios (e.g., [21]). In the reverse matching, we used the features extracted through VggNet [17]. Subsequently, based on the calculated distances, in each iteration we record the position of the original test image among the ranked list. Thus, if the representative set (acting as a new test object), and the test image (acting as a training sample) belong to the same class, the test image shall score the lowest distance (top in the ranked list), thus gaining the highest score. Finally, for each test image, this strategy results in a *1xM* vector containing reverse matching scores of each test image with respect to the representative set from each event category. A step wise elaboration of the reverse matching strategy is provided in Algorithm 2.

Ultimately the final decision of our system finalizes by multiplying the IOWA classification scores by the reverse matching scores (i.e., the inverse of its position index) in order to draw a final classification score, and the highest value therein defines the class to which the test image belongs.

3. EXPERIMENTS AND RESULTS

3.1. Dataset

To validate our framework, we employ a recently-introduced large-scale benchmark dataset known as "Web Images Dataset for Event Recognition (WIDER)" [22]. WIDER holds a total of around 60,000 images from 61 different event classes. It stems as the largest and most complex benchmark for event recognition in still images to date. It covers diverse event classes, where most of the event classes are taken from Large Scale Ontology for Multimedia (LSCOM) [23]. For instance, it contains event categories from sports (such as football, basketball and tennis), daily life events (such as shopping and meeting) and social events (such as concert, celebration and funeral). Moreover, it also covers some specific events, such as demonstration, riots, surgery and soldier marching and drilling. It is a very convenient choice to evaluate our framework, thanks to its size, diversity, and challenging nature.

3.2. Experimental Configurations

To assess the performance of different network architectures, and the importance of scene and object level-information, we conducted 3 different experiments, as follows:

Experiment # 1: We investigate the performance of individual pre-trained networks in event recognition. Here, features are extracted from each single pre-trained network.

Experiment # 2: We judge the performance of IOWA-fused networks, of the same as well as different architectures.

Experiment # 3: We include the reverse matching stage to raise the classification rates.

3.3. Experimental Results

Table 1 reports the results from our first experiment. As can be seen, VggNet [17] yielded higher scores than the remaining architectures, particularly when pre-trained on ImageNet dataset [18]. Another interesting observation is that all architectures show better results in event recognition when pre-trained on ImageNet dataset, which suggests that object information seem to be more pivotal for the task of event recognition in static images. However, the opposite behavior has been noticed for some classes, such as concert, spa and press conference. Moreover, though overall VGGNet achitecture performs better, we observed better performances for other architectures on some events. For instance, on picnic images,

Table 1. Classification results with individual networks

Network	Avg. Acc.
VGGNet ImageNet Features	45.78%
VGGNet Places Features	43.46%
AlexNet ImageNet Features	39.83%
AlexNet Places Features	39.71%
GoogleNet ImageNet Features	42.78%
GoogleNet Places Features	40.05%

Table 2. Classification results of different combinations with the fusion strategy

Network	Avg. Acc.
VGGNet (O+S)	54.22%
AlexNet (O+S)	49.28%
GoogleNet (O+S)	51.00%
GoogleNet (O) + AlexNet (O)	52.00%
VGGNet (O) + AlexNet (O)	53.37%
VGGNet (O) + GoogleNet (O)	53.80%
VGGNet (S) + AlexNet (S)	52.93%
VGGNet (O+S) + AlexNet (O+S)	56.86%
VGGNet (O+S) + GoogleNet (O+S)	57.02%
VGGNet (O+S) + AlexNet (O+S) + GoogleNet(O+S)	58.05%
VGGNet (O+S) + AlexNet (O+S) + GNet(O+S)+ B-Matching	59.49%

Alexnet [15] has better results compared to the other competitors (page limitations restrict the display of per-class accuracies). This variation in the performances of network architectures was a key-inspiration for the fusion of classification results from different network architectures. In Table 2, we investigate the performance of different combinations of pre-trained networks via the IOWA fusion. The gain incurred by the fusion stage can be easily spotted. Initially, we investigate different network pairs. Particularly, we are interested, as noted before, in different combinations of scene and object information extracted through different architectures. In Table 2, slight variation in the performances can be noticed, where the combination of classifiers trained on places and object level features, extracted through VGGNet, provides better results. It supports our intuition that both objects and places information, if well-capitalized on, can indeed boost the results.

Another interesting observation is that the combination of different architectures pre-trained on object dataset yields better results compared to the ones pre-trained on places dataset. However, the combined results of all architectures (i.e., VGGNet, GoogleNet and AlexNet) pre-trained on both datasets are significantly higher than the pair combinations. Thus, this confirms that combining the classification scores from multiple architectures improves the overall results.

In our third experiment, the reverse matching stage is included. The reverse matching strategy improves the overall result by 1.44%, which is still interesting with regards to the challenging nature of the employed dataset. This also suggests that reverse matching can further improve the results in event recognition context.

We also compare our method with recent state-of-the-art on three challenging datasets in Table 3 and Table 4. Table 3 provides the comparisons results on WIDER and UIUC datasets while the comparisons against state of the art on USED dataset [9] are shown in Table 4. In [22], features from different layers have been combined for event recognition. Similarly, in [10], different techniques have been proposed for object-scene transferring and the final classification decision represents the average score of individual image regions. On the WIDER dataset, our approach has an overall gain of 17% and 6.4% over the deep fusion method [22] and the best method in [10], respectively.

Table 3. Comparisons against state of the art on WIDER and UIUC Datasets

WIDER Dataset		**UIUC Dataset**	
Method	Avg. Acc.	Method	Avg. Acc.
Baseline Method [22]	39.7%	Baseline Method [24]	73.40%
Deep Channel Fusion [22]	42.4%	Places CNN Features [19]	94.10%
Method in [25]	44.06%	GoogleNet GAP [26]	95.00%
Method in [10]	53.0%	Method in [10]	98.80%
Our Approach	**59.49%**	**Our Approach**	**99.22%**

Similarly, on UIUC dataset, which mainly covers sports events, and USED dataset, which contains images from 14 different social events, our approach exhibit plausible performances with respect to recent leading works on both datasets. The main strength of our approach is that it assigns fusion weights to each classifier based on its confidence, which leads to a better utilization of object/scene information extracted through different CNN architectures.

Table 4. Comparisons against state of the art on USED [9]

Methods	Avg. Acc. on Subset 1	Avg. Acc. on Subset 2
Baseline Method [9]	70.03%	65.96%
Method [27]	72%	79.29%
Our Approach	**79.13%**	**87.02%**

4. CONCLUSION

This paper proposes a novel framework for event recognition in still images. We show that our framework, which couples an IOWA fusion scheme with reverse matching scores, yields very good results with respect to recent works on three benchmark datasets. Nevertheless, better results can be achieved by considering for instance (i) feature learning, in order to further improve the representativeness of the features, and (ii) adopting some saliency-related features given that, in an event image, not all the image regions hold valuable visual information in characterizing the image in hand but only the salient ones do. We believe that investing in these two directions can expectedly improve the overall event recognition accuracy.

5. REFERENCES

[1] C. Tzelepis, Z. Ma, V. Mezaris, B. Ionescu, I. Kompatsiaris, G. Boato, N. Sebe, and S. Yan, "Event-based media processing and analysis: A survey of the literature," *Image and Vision Computing*, 2016.

[2] R. Mattivi, G. Boato, and F. G. B. De Natale, "Event-based media organization and indexing," *Infocommunications Journal*, vol. 3, no. 3, pp. 9–18, 2011.

[3] A. Rosani, G. Boato, and F. G. B. De Natale, "Eventmask: A game-based framework for event-saliency identification in images," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1359–1371, 2015.

[4] A. Salvador, M. Zeppelzauer, D. Manchon-Vizuete, A. Calafell, and X. Giro-i Nieto, "Cultural event recognition with visual convnets and temporal models," in *CVPR Workshops*, 2015, pp. 36–44.

[5] L. Wang, Z. Wang, W. Du, and Y. Qiao, "Object-scene convolutional neural networks for event recognition in images," in *CVPR Workshops*, 2015, pp. 30–35.

[6] C. Gan, N. Wang, Y. Yang, D. Yeung, and A. G Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *CVPR*, 2015, pp. 2568–2577.

[7] S. Park and N. Kwak, "Cultural event recognition by subregion classification with convolutional neural network," in *CVPR Workshops*, 2015, pp. 45–50.

[8] Kashif Ahmad, Francesco De Natale, Giulia Boato, and Andrea Rosani, "A hierarchical approach to event discovery from single images using mil framework," in *Signal and Information Processing (GlobalSIP), 2016 IEEE Global Conference on*. IEEE, 2016, pp. 1223–1227.

[9] K. Ahmad, N. Conci, G. Boato, and F. G. B. De Natale, "Used: a large-scale social event detection dataset," in *ACM Multimedia Systems*, 2016, pp. 50–55.

[10] L. Wang, Z. Wang, Y. Qiao, and L. V. Gool, "Transferring object-scene convolutional neural networks for event recognition in still images," *arXiv preprint arXiv:1609.00162*, 2016.

[11] L. Wang, Z. Wang, S. Guo, and Y. Qiao, "Better exploiting os-cnns for better event recognition in images," in *CVPR Workshops*, 2015, pp. 45–52.

[12] R. R Yager and D. P Filev, "Induced ordered weighted averaging operators," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 29, no. 2, pp. 141–150, 1999.

[13] Q. Leng, R. Hu, C. Liang, Y. Wang, and J. Chen, "Bidirectional ranking for person re-identification," in *IEEE ICME*, 2013, pp. 1–6.

[14] Z. Wu, Q. Ke, J. Sun, and H. Shum, "Scalable face image retrieval with identity-based quantization and multireference reranking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 10, pp. 1991–2001, 2011.

[15] A. Krizhevsky, I. Sutskever, and G. E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[19] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014, pp. 487–495.

[20] Y. Bazi, N. Alajlan, Farid Melgani, et al., "Robust estimation of water chlorophyll concentrations with gaussian process regression and iowa aggregation operators," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 7, pp. 3019–3028, 2014.

[21] J. Wang and J. Zucker, "Solving multiple-instance problem: A lazy learning approach," 2000.

[22] Y. Xiong, K. Zhu, D. Lin, and X. Tang, "Recognize complex events from static images by fusing deep channels," in *CVPR*, 2015, pp. 1600–1609.

[23] M. Naphade, J. R Smith, J. Tesic, S. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[24] L. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *ICCV*, 2007, pp. 1–8.

[25] R. F Rachmadi, K. Uchimura, and G. Koutaki, "Combined convolutional neural network for event recognition," *Presented in FCV 2016*.

[26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE CVPR*, 2016, pp. 2921–2929.

[27] R. F. Rachmadi, K. Uchimura, and G. Koutaki, "Spatial pyramid convolutional neural network for social event detection in static image," *arXiv preprint arXiv:1612.04062(presented in ICAST 2016)*, 2016.