Open access • Proceedings Article • DOI:10.1109/CVPRW.2015.7301329

ChaLearn Looking at People 2015 challenges: Action spotting and cultural event recognition — Source link

Xavier Baró, Jordi Gonzàlez, Junior Fabian, Miguel Ángel Bautista ...+4 more authors

Institutions: Open University of Catalonia, University of Barcelona

Published on: 07 Jun 2015 - Computer Vision and Pattern Recognition

ChaLearn Looking at People 2015 challenges: action spotting and cultural event recognition

Xavier Baró
Universitat Oberta de Catalunya - CVC
xbaro@uoc.edu

Jordi Gonzàlez
Computer Vision Center (UAB)
poal@cvc.uab.es

Junior Fabian
Computer Vision Center (UAB)
jfabian@cvc.uab.es

Miguel A. Bautista
Universitat de Barcelona
miguelangelbautistamartin@gmail.com

Marc Oliu
Universitat de Barcelona
moliusimon@gmail.com

Hugo Jair Escalante
INAOE
hugo.jair@gmail.com

Isabelle Guyon
Chalearn
guyon@chalearn.org

Sergio Escalera
Universitat de Barcelona - CVC
sergio@maia.ub.es

Abstract

Following previous series on Looking at People (LAP) challenges [6, 5, 4], ChaLearn ran two competitions to be presented at CVPR 2015: action/interaction spotting and cultural event recognition in RGB data. We ran a second round on human activity recognition on RGB data sequences. In terms of cultural event recognition, tens of categories have to be recognized. This involves scene understanding and human analysis. This paper summarizes the two challenges and the obtained results. Details of the ChaLearn LAP competitions can be found at http://gesture.chalearn.org/.

1. Introduction

The automatic analysis of the human body in still images and image sequences, also known as Looking at People, keeps making rapid progress with the constant improvement of new published methods that push the state-of-the-art. Applications are countless, like Human Computer Interaction, Human Robot Interaction, communication, entertainment, security, commerce and sports, while having an important social impact in assistive technologies for the handicapped and the elderly.

In 2015, ChaLearn[1] organized new competitions and CVPR workshop on action/interaction spotting and cultural event recognition. The recognition of continuous, natural human signals and activities is very challenging due to the multimodal nature of the visual cues (e.g., movements of fingers and lips, facial expression, body pose), as well as technical limitations such as spatial and temporal resolution. In addition, images of cultural events constitute a very challenging recognition problem due to a high variability of garments, objects, human poses and context. Therefore, how to combine and exploit all this knowledge from pixels constitutes a challenging problem.

This motivates our choice to organize a new workshop and a competition on this topic to sustain the effort of the computer vision community. These new competitions come as a natural evolution from our previous workshops at CVPR 2011, CVPR 2012, ICPR 2012, ICMI 2013, and ECCV 2014. We continued using our website http://gesture.chalearn.org for promotion, while challenge entries in the quantitative competition were scored on-line using the Codalab Microsoft-Stanford University platforms (http://codalab.org/), from which we have already organized international challenges related to Computer Vision and Machine Learning problems.

In the rest of this paper, we describe in more detail the organized challenges and obtained results by the participants of the competition.

[1] www.chalearn.org

2. Challenge tracks and schedule

The ChaLearn LAP 2015 challenge featured two quantitative evaluations: action/interaction spotting on RGB data and cultural event recognition in still images. The characteristics of both competition tracks are the following:

- Action/Interaction recognition: in total, 235 action samples performed by 17 actors were provided. The selected actions involved the motion of most of the limbs and included interactions among various actors.
- Cultural event recognition: Inspired by the Action Classification challenge of PASCAL VOC 2011-12 successfully organized by Everingham*et al*. [8], we planned to run a competition in which 50 categories corresponding to different world-wide cultural events would be considered. In all the image categories, garments, human poses, objects, illumination, and context do constitute the possible cues to be exploited for recognizing the events, while preserving the inherent inter- and intra-class variability of this type of images. Thousands of images were downloaded and manually labeled, corresponding to cultural events like Carnival (Brasil, Italy, USA), Oktoberfest (Germany), San Fermin (Spain), Holi Festival (India) and Gion Matsuri (Japan), among others.

The challenge was managed using the Microsoft Codalab platform[2]. The schedule of the competition was as follows.

December 1st, 2014 Beginning of the quantitative competition for action/interaction recognition track, release of development and validation data.

January 2nd, 2015 Beginning of the quantitative competition for cultural event recognition track, release of development and validation data.

February 15th, 2015: Beginning of the registration procedure for accessing to the final evaluation data.

March 13th, 2015: Release of the encrypted final evaluation data and validation labels. Participants started training their methods with the whole dataset.

March 13th, 2015: Release of the decryption key for the final evaluation data. Participants started predicting the results on the final evaluation labels. This date was the deadline for code submission as well.

March 20th, 2015: End of the quantitative competition. Deadline for submitting the predictions over the final evaluation data. The organizers started the code verification by running it on the final evaluation data.

March 25th, 2015: Deadline for submitting the fact sheets.

March 27th, 2015: Publication of the competition results.

[2]https://www.codalab.org/competitions/

3. Competition data

This section describes the datasets provided for each competition and its main characteristics.

3.1. Action and Interaction dataset

We provided the *HuPBA* $8K+$ dataset [15] with annotated begin and end frames of actions and interactions. A key frame example for each action/interaction category is shown in Figure 1. The characteristics of the dataset are:

- The images are obtained from 9 videos (RGB sequences) and a total of 14 different actors appear in the sequences. The image sequences have been recorded using a stationary camera with the same static background.
- 235 action/interaction samples performed by 14 actors.
- Each video (RGB sequence) was recorded at 15 fps rate, and each RGB image was stored with resolution 480×360 in BMP file format.
- 11 action categories, containing isolated and collaborative actions: Wave, Point, Clap, Crouch, Jump, Walk, Run, Shake Hands, Hug, Kiss, Fight. There is a high intra-class variability among action samples.
- The actors appear in a wide range of different poses and performing different actions/gestures which vary the visual appearance of human limbs. So there is a large variability of human poses, self-occlusions and many variations in clothing and skin color.
- Large difference in length about the performed actions and interactions. Several distractor actions out of the 11 categories are also present.

A list of data attributes for this track dataset is described in Table 1. Examples of images of the data set are shown in Figure 1.

3.2. Cultural Event Recognition dataset

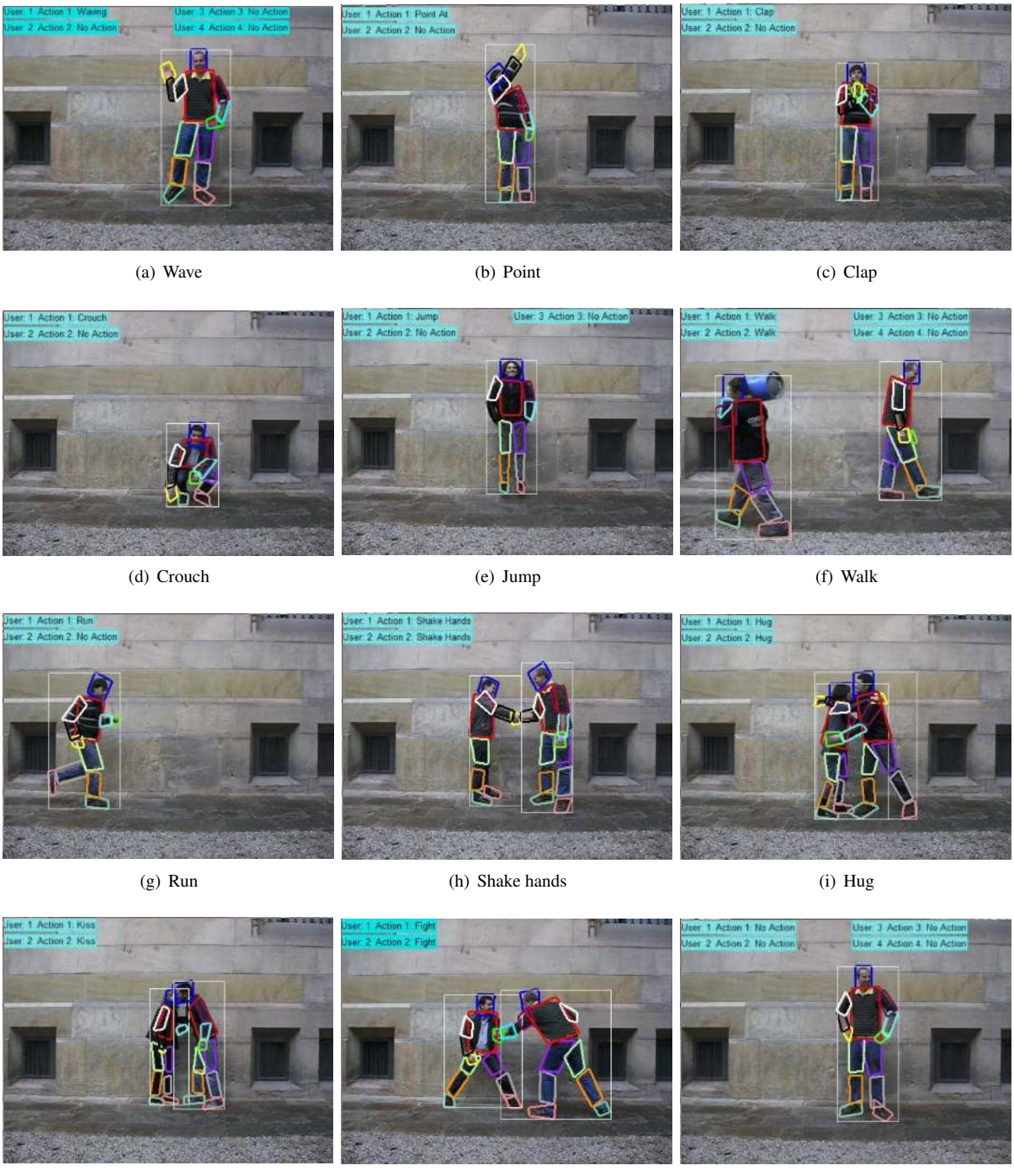
In this work, we introduce the first dataset based on cultural events and the first cultural event recognition challenge. In this section, we discuss some of the works most closely related to it.

Action Classification Challenge [8] This challenge belongs to the PASCAL - VOC challenge which is a benchmark in visual object category recognition and detection. In particular, the Action Classification challenge was introduced in 2010 with 10 categories. This challenge consisted on predicting the action(s) being performed by a person in a still image. In 2012 there were two variations of this competition, depending on how the person (whose actions are to be classified) was identified in a test image: (i) by a tight

Training actions	Validation actions	Test actions	Sequence duration	FPS
150	90	95	9× 1-2 min	15

Modalities	Num. of users	Action categories	interaction categories	Labeled sequences
RGB	14	7	4	235

Table 1. Action and interaction data characteristics.



(a) Wave (b) Point (c) Clap

(d) Crouch (e) Jump (f) Walk

(g) Run (h) Shake hands (i) Hug

(j) Kiss (k) Fight (l) Idle

Figure 1. Key frames of the *HuPBA* $8K+$ dataset used in the action/interaction recognition track, showing actions ((a) to (g)), interactions ((h) to (k)) and the idle pose (l).

Dataset	#Images	#Classes	Year
Action Classification Dataset [8]	5,023	10	2010
Social Event Dataset [11]	160,000	149	2012
Event Identification Dataset [1]	594,000	24,900	2010
Cultural Event Dataset	11,776	50	2015

Table 2. Comparison between our cultural event dataset and others present in the state of the art.

bounding box around the person; (ii) by only a single point located somewhere on the body.

Social Event Detection [11] This work is composed of three challenges and a common test dataset of images with their metadata (timestamps,tags, geotags for a small subset of them). The first challenge consists of finding technical events that took place in Germany in the test collection. In the second challenge, the task consists of finding all soccer events taking place in Hamburg (Germany) and Madrid (Spain) in the test collection. The third challenge aims at finding demonstration and protest events of the *Indignados* movement occurring in public places in Madrid in the test collection.

Event Identification in Social Media [1] In this work the authors introduce the problem of event identification in social media. They presented an incremental clustering algorithm that classifies social media documents into a growing set of events.

Table 2 shows a comparison between our cultural event dataset and the others present in the state of the art. Action Classification dataset is the most closely related, but the amount of images and categories is smaller than ours. Although the number of images and categories in the datasets [11] and [1] are larger than our dataset, these dataset are not related to cultural events but to events in general. Some examples of the events considered in these dataset are soccer events (*football games that took place in Rome in January 2010*), protest events (*Indignados movement occurring in public places in Madrid*), etc.

3.2.1 Dataset

The Cultural Event Recognition challenge aims to investigate the performance of recognition methods based on several cues like garments, human poses, objects, background, etc. To this end, the cultural event dataset contains significant variability in terms of clothes, actions, illumination, localization and context.

The Cultural Event Recognition dataset consists of images collected from two image search engines (Google Images and Bing Images). To build the dataset, we chose 50 important cultural events in the world and we created several queries with the names of these events. In order to increase the number of retrieved images, we combined the names of the events with some additional keywords (festival, parade, event, etc.). Then, we removed duplicated URLs and downloaded the raw images. To ensure that the downloaded images belonged to each cultural event, a process was applied to manually filter each of the images. Next, all exact duplicate and near duplicate images were removed from the downloaded image set using the method described in[3]. While we attempted to remove all duplicates from the dataset, there may exist some remaining duplicates that were not found. We believe the number of these is small enough so that they will not significantly impact research. After all this preprocessing, our dataset is composed of $11,776$ images. Figure 2 depicts in shades of green the amount of cultural events selected by country.

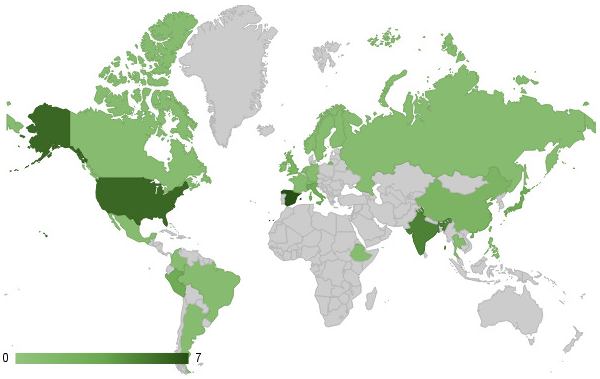


Figure 2. Cultural events by country, dark green represents greater number of events.

The dataset can be viewed and downloaded at the following web address: https://www.codalab.org/competitions/2611. Some additional details and main contributions of the cultural event dataset are described below:

- First dataset on cultural events from all around the globe.
- More than $11,000$ images representing 50 different categories.
- High intra- and inter-class variability.
- For this type of images, different cues can be exploited like garments, human poses, crowds analysis, objects and background scene.
- The evaluation metric will be the recognition accuracy.

Figure 3 shows some sample images and Table 3 lists the 50 selected cultural events, country they belong and the number of images considered for this challenge.

There is no similar dataset in the literature. For example, the ImageNet competition does not include the cultural event taxonomy as considered in this specific track. Considering the Action Classification challenge of PASCAL VOC

Figure 3. Cultural events sample images.

2011-12, the number of images is similar, around $11,000$, but the number of categories is here increased more than 5 times.

4. Protocol and evaluation

This section introduces the protocol and evaluation metrics for both tracks.

4.1. Evaluation procedure for action/interaction track

To evaluate the accuracy of action/interaction recognition, we use the Jaccard Index, the higher the better. Thus, for the n action and interaction categories labeled for a RGB sequence s, the Jaccard Index is defined as:

$$J_{s,n} = \frac{A_{s,n} \bigcap B_{s,n}}{A_{s,n} \bigcup B_{s,n}}, \quad (1)$$

where $A_{s,n}$ is the ground truth of action/interaction n at sequence s, and $B_{s,n}$ is the prediction for such an action at sequence s. $A_{s,n}$ and $B_{s,n}$ are binary vectors where 1-values correspond to frames in which the $n-$th action is being performed. The participants were evaluated based on the mean Jaccard Index among all categories for all sequences, where motion categories are independent but not mutually exclusive (in a certain frame more than one action, interaction, gesture class can be active).

In the case of false positives (e.g. inferring an action

Cultural Event	Country	#Images
1. Annual Buffalo Roundup	USA	334
2. Ati-atihan	Philippines	357
3. Ballon Fiesta	USA	382
4. Basel Fasnacht	Switzerland	310
5. Boston Marathon	USA	271
6. Bud Billiken	USA	335
7. Buenos Aires Tango Festival	Argentina	261
8. Carnival of Dunkerque	France	389
9. Carnival of Venice	Italy	455
10. Carnival of Rio	Brazil	419
11. Castellers	Spain	536
12. Chinese New Year	China	296
13. Correfocs	Catalonia	551
14. Desert Festival of Jaisalmer	India	298
15. Desfile de Silleteros	Colombia	286
16. Día de los Muertos	Mexico	298
17. Diada de Sant Jordi	Catalonia	299
18. Diwali Festival of Lights	India	361
19. Falles	Spain	649
20. Festa del Renaixement Tortosa	Catalonia	299
21. Festival de la Marinera	Peru	478
22. Festival of the Sun	Peru	514
23. Fiesta de la Candelaria	Peru	300
24. Gion matsuri	Japan	282
25. Harbin Ice and Snow Festival	China	415
26. Heiva	Tahiti	286
27. Helsinki Samba Carnival	Finland	257
28. Holi Festival	India	553
29. Infiorata di Genzano	Italy	354
30. La Tomatina	Spain	349
31. Lewes Bonfire	England	267
32. Macys Thanksgiving	USA	335
33. Maslenitsa	Russia	271
34. Midsommar	Sweden	323
35. Notting hill carnival	England	383
36. Obon Festival	Japan	304
37. Oktoberfest	Germany	509
38. Onbashira Festival	Japan	247
39. Pingxi Lantern Festival	Taiwan	253
40. Pushkar Camel Festival	India	433
41. Quebec Winter Carnival	Canada	329
42. Queens Day	Netherlands	316
43. Rath Yatra	India	369
44. SandFest	USA	237
45. San Fermin	Spain	418
46. Songkran Water Festival	Thailand	398
47. St Patrick's Day	Ireland	320
48. The Battle of the Oranges	Italy	276
49. Timkat	Ethiopia	425
50. Viking Festival	Norway	262

Table 3. List of the 50 Cultural Events.

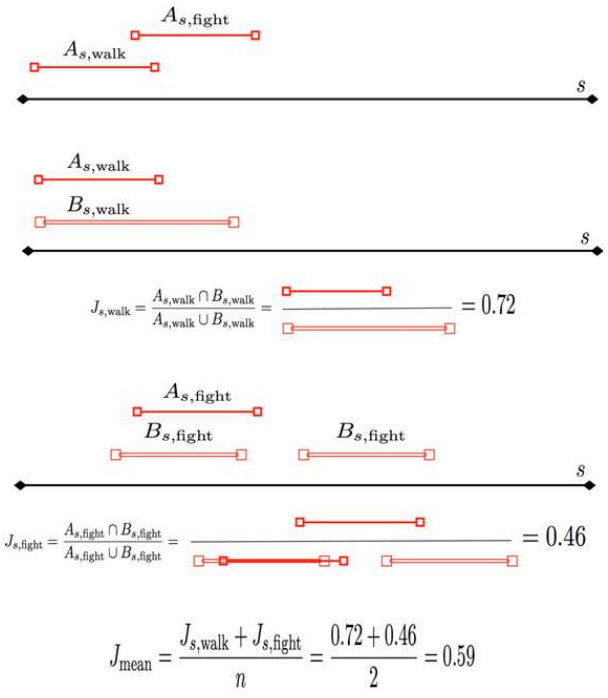


Figure 4. Example of mean Jaccard Index calculation.

or interaction not labeled in the ground truth), the Jaccard Index is 0 for that particular prediction, and it will not count in the mean Jaccard Index computation. In other words n is equal to the intersection of action/interaction categories appearing in the ground truth and in the predictions.

An example of the calculation for two actions is shown in Figure 4. Note that in the case of recognition, the ground truth annotations of different categories can overlap (appear at the same time within the sequence). Also, although different actors appear within the sequence at the same time, actions/interactions are labeled in the corresponding periods of time (that may overlap), there is no need to identify the actors in the scene.

The example in Figure 4 shows the mean Jaccard Index calculation for different instances of actions categories in a sequence (single red lines denote ground truth annotations and double red lines denote predictions). In the top part of the image one can see the ground truth annotations for actions walk and fight at sequence s. In the center part of the image a prediction is evaluated obtaining a Jaccard Index of 0.72. In the bottom part of the image the same procedure is performed with the action fight and the obtained Jaccard Index is 0.46. Finally, the mean Jaccard Index is computed obtaining a value of 0.59.

4.2. Evaluation procedure for cultural event track

For the cultural event track, participants were asked to submit for each image their confidence for each of the events. Participants submissions were evaluated using the average precision (AP), inspired in the metric used for PASCAL challenges [7]. It is calculated as follows:

1. First, we compute a version of the precision/recall curve with precision monotonically decreasing. It is obtained by setting the precision for recall r to the maximum precision obtained for any recall $r' \geq r$.

2. Then, we compute the AP as the area under this curve by numerical integration. For this, we use the well-know trapezoidal rule. Let $f(x)$ the function that represents our precision/recall curve, the trapezoidal rule works by approximating the region under this curve as follows:

$$\int_a^b f(x)dx \approx (b-a)\left[\frac{f(a)+f(b)}{2}\right] \qquad (2)$$

5. Challenge results and methods

In this section we summarize the methods proposed by the top ranked participants. Eight teams submitted their code and predictions for the last phase of the competition, two for action/interaction and six for cultural event. Table 4 contains the final team rank and score for both tracks, and the methods used for each team are described in the rest of this section.

5.1. Action/Interaction recognition methods

MMLAB: This method is an improvement of the system proposed in [13], which is composed of two parts: video representation and temporal segmentation. For the representation of video clip, the authors first extracted improved dense trajectories with HOG, HOF, MBHx, and MBHy descriptors. Then, for each kind of descriptor, the participants trained a GMM and used Fisher vector to transform these descriptors into a high dimensional super vector space. Finally, sum pooling was used to aggregate these codes in the whole video clip and normalize them with power L2 norm. For the temporal recognition, the authors resorted to a temporal sliding method along the time dimension. To speed up the processing of detection, the authors designed a temporal integration histogram of Fisher Vector, with which the pooled Fisher Vector was efficiently evaluated at any temporal window. For each sliding window, the authors used the pooled Fisher Vector as representation and fed it into the SVM classifier for action recognition. A summary of this method is shown in Figure 5.

FKIE: The method implements an end-to-end generative approach from feature modeling to activity recognition. The system combines dense trajectories and

Action/Interaction Track								
Rank	Team name	Score	Features	Dimension reduction	Clustering	Classification	Temporal coherence	Action representation
1	MMLAB	**0.5385**	IDT [19]	PCA	-	SVM	-	Fisher Vector
2	FIKIE	**0.5239**	IDT	PCA	-	HMM	Appearance+Kalman filter	-
Cultural Event Track								
Rank	Team name	Score	Features			Classification		
1	MMLAB	**0.855**	Multiple CNN			Late weighted fusion of CNNs predictions.		
2	UPC-ST	**0.767**	Multiple CNN			SVM and late weighted fusion.		
3	MIPAL_SNU	**0.735**	Discriminant regions [18] + CNNs			Entropy + Mean Probabilities of all patches		
4	SBU_CS	**0.610**	CNN-M [2]			SPM [10] based on LSSVM [16]		
5	MasterBlaster	**0.58**	CNN			SVM, KNN, LR and One Vs Rest		
6	Nyx	**0.319**	Selective-search approach [17] + CNN			Late fusion AdaBoost		

Table 4. Chalearn LAP 2015 results.

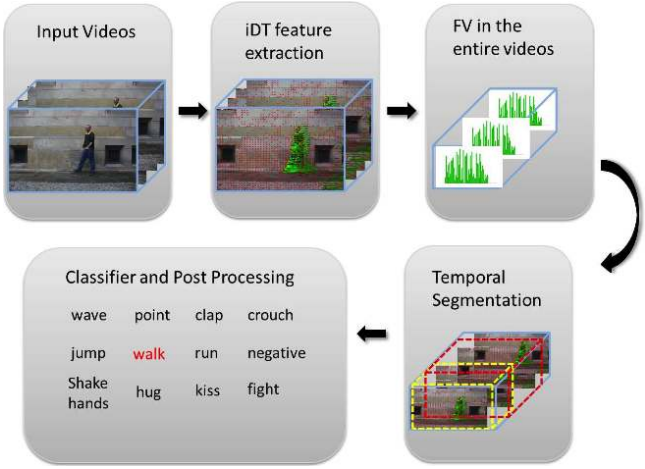
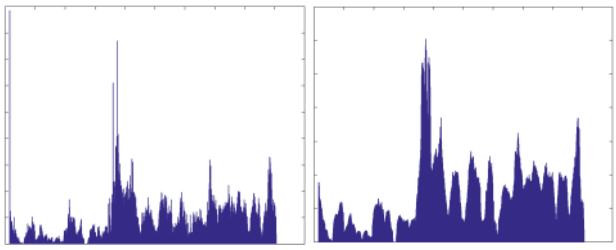


Figure 5. Method summary for MMLAB team [21].



(a) (b)

Figure 6. Example of DT feature distribution for the first 200 frames of Seq01 for FKIE team, (a) shows the distribution of the original implementation, (b) shows the distribution of their version.

Fisher Vectors with a temporally structured model for action recognition based on a simple grammar over action units. The authors modify the original dense trajectory implementation of Wang*et al*. [19] to avoid the omission of neighborhood interest points once a trajectory is used (the improvement is shown in Figure 6). They use an open source speech recognition engine for the parsing and segmentation of video sequences. Because a large data corpus is typically needed for training such systems, images were mirrored to artificially generate more training data. The final result is achieved by voting over the output of various parameter and grammar configurations.

5.2. Cultural event recognition methods

MMLAB: This method fuses five kinds of ConvNets for event recognition. Specifically, they fine-tune Clarifai net pre-trained on the ImageNet dataset, Alex net pre-trained on Places dataset, Googlenet pre-trained on the ImageNet dataset and the Places dataset, and VGG 19-layer net on the ImageNet dataset. The prediction scores from these five ConvNets are weighted fused as final results. A summary of this method is shown in Figure 7.

UPC-STP: This solution was based on combining the features from the fully connected (FC) layers of two convolutional neural networks (ConvNets): one pre-trained with ImageNet images and a second one fine-tuned with the ChaLearn Cultural Event Recognition dataset. A linear SVM was trained for each of the features associated to each FC layer and later fused with an additional SVM classifier, resulting into a hierarchical architecture. Finally the authors refined their solution by weighting the outputs of the FC classifiers with a temporal modeling of the events learned from the training data. In particular, high classification scores based on visual features were penalized when their time stamp did not match well an event-specific temporal distribution. A summary of this method is shown in Figure 8.

MIPAL_SNU: The motivation of this method is that training and testing with only the discriminant regions will improve the performance of classification. Inspired by [9], they first extract region proposals which are candidates of the distinctive regions for cultural event recognition. Work [18] was used to detect possibly meaningful regions of various size. Then, the patches are trained using deep convolutional neural network (CNN) which has 3 convolutional layers and pooling layers, and 2 fully-connected networks. After training, probability distribution for the classes is calculated for every image patch from test image. Then, class proba-

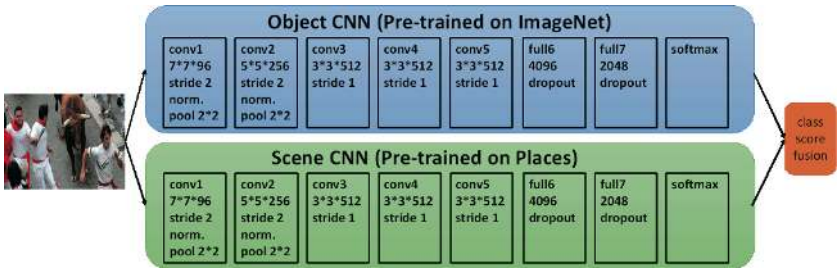


Figure 7. Method summary for MMLAB team [20].

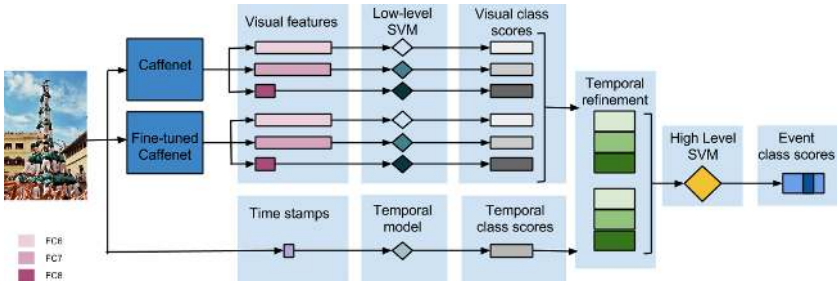


Figure 8. Method summary for UPC-STP team [14].

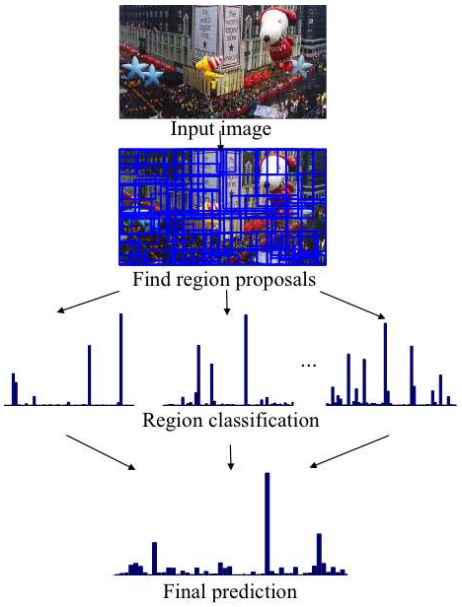


Figure 9. Method summary for MIPAL_SNU team [12].

bilities of the test image is determined as a mean of the probabilities of all patches after the entropy thresholding step.

6. Discussion

This paper described the main characteristics of the ChaLearn Looking at People 2015 Challenge which included competitions on (i) RGB action/interaction recognition and (ii) cultural event recognition. Two large datasets were designed, manually-labelled, and made publicly available to the participants for a fair comparison in the performance results. Analysing the methods used by the teams that finally participated in the test set and uploaded their models, several conclusions can be drawn.

For the case of action and interaction recognition in RGB data sequences, all the teams used Improved Dense Trajectories [19] as features, using PCA for dimensionality reduction. From the point of view of the classifiers, both generative and discriminative have been used by teams, although SVM obtained better results. This is the second round of this competition and the proposed methods outperform the ones from the first round. Nevertheless, since on the development phase of the competition only the two finalists obtained better results than the baseline and the winner score has been of 0.5385, it denotes that there is still room for improvement, and that action/interaction recognition is still an open problem.

In the case of cultural event recognition, and following current trends in the computer vision literature, deep learning architecture is present in most of the solutions. Since the huge number of images required for training Convolutional Neural Networks, teams used standard pre-trained networks as input to their systems, followed by different types of classification strategies.

The complexity and computational requirements of some of the state of the art methods made them unfeasible for this kind of competitions where time is a hard restriction. However, the irruption of GPU computation on research, that has been used by many teams in both tracks, has enabled those methods, with a great impact on the final results.

Acknowledgements

We would like to thank to the sponsors of these competitions: Microsoft Research, University of Barcelona, Amazon, INAOE, VISADA, and California Naturel. This research has been partially supported by research projects TIN2012.38187-C02-02, TIN2012-39051 and TIN2013-43478-P. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for creating the baseline of the Cultural Event Recognition track.

References

[1] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in socialmedia. In *Proceedings WSDM*, 2010.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *CoRR*, abs/1405.3531, 2014.

[3] O. Chum, J. Philbin, M. Isard, and A. Zisserman. Scalable near identical image and shot detection. In *ACM International Conference on Image and Video Retrieval*, 2007.

[4] S. Escalera, X. Baró, J. Gonzàlez, M. Bautista, M. Madadi, M. Reyes, V. Ponce, H. Escalante, J. Shotton, and I. Guyon. Chalearn looking at people challenge 2014: Dataset and results. *ChaLearn Looking at People, European Conference on Computer Vision*, 2014.

[5] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, I. Guyon, V. Athitsos, H. Escalante, A. Argyros, C. Sminchisescu, R. Bowden, and S. Sclarof. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. *15th ACM International Conference on Multimodal Interaction*, pages 365–368, 2013.

[6] S. Escalera, J. Gonzàlez, X. Baró, M. Reyes, O. Lopés, I. Guyon, V. Athitsos, and H. J. Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *ChaLearn Multi-Modal Gesture Recognition Grand Challenge and Workshop, 15th ACM International Conference on Multimodal Interaction*, 2013.

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

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[9] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE, 2014.

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

[11] S. Papadopoulos, E. Schinas, V. Mezaris, R. Troncy, and I. Kompatsiaris. Social event detection at mediaeval 2012: Challenges, dataset and evaluation. In *Proc. MediaEval 2012 Workshop*, 2012.

[12] S. Park and N. Kwak. Cultural event recognition by subregion classification with convolutional neural network. In *In CVPR ChaLearn Looking at People Workshop 2015*, 2015.

[13] X. Peng, L. Wang, Z. Cai, and Y. Qiao. Action and gesture temporal spotting with super vector representation. In L. Agapito, M. M. Bronstein, and C. Rother, editors, *Computer Vision - ECCV 2014 Workshops*, volume 8925 of *Lecture Notes in Computer Science*, pages 518–527. Springer International Publishing, 2015.

[14] A. Salvador, M. Zeppelzauer, D. Monchon-Vizuete, A. Calafell, and X. Giro-Nieto. Cultural event recognition with visual convnets and temporal models. In *In CVPR ChaLearn Looking at People Workshop 2015*, 2015.

[15] D. Sánchez, M. A. Bautista, and S. Escalera. HuPBA 8k+: Dataset and ECOC-graphcut based segmentation of human limbs. *Neurocomputing*, 2014.

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

[17] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[18] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[19] H. Wang and C. Schmid. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*, 2013.

[20] L. Wang, Z. Wang, W. Du, and Q. Yu. Event recognition using object-scene convolutional neural networks. In *In CVPR ChaLearn Looking at People Workshop 2015*, 2015.

[21] Z. Wang, L. Wang, W. Du, and Q. Yu. Action spotting system using fisher vector. In *In CVPR ChaLearn Looking at People Workshop 2015*, 2015.